MRC | Epidemiology Unit

Analysis of Complex Traits I: Overview and Case Studies

Jing Hua Zhao

MRC Epidemiology Unit, Cambridge, UK
jinghua.zhao@mrc-epid.cam.ac.uk
http://www.mrc-epid.cam.ac.uk/~jinghua.zhao

11 August 2008, Dortmund, Germany

useR! 2008

Outline of Contents

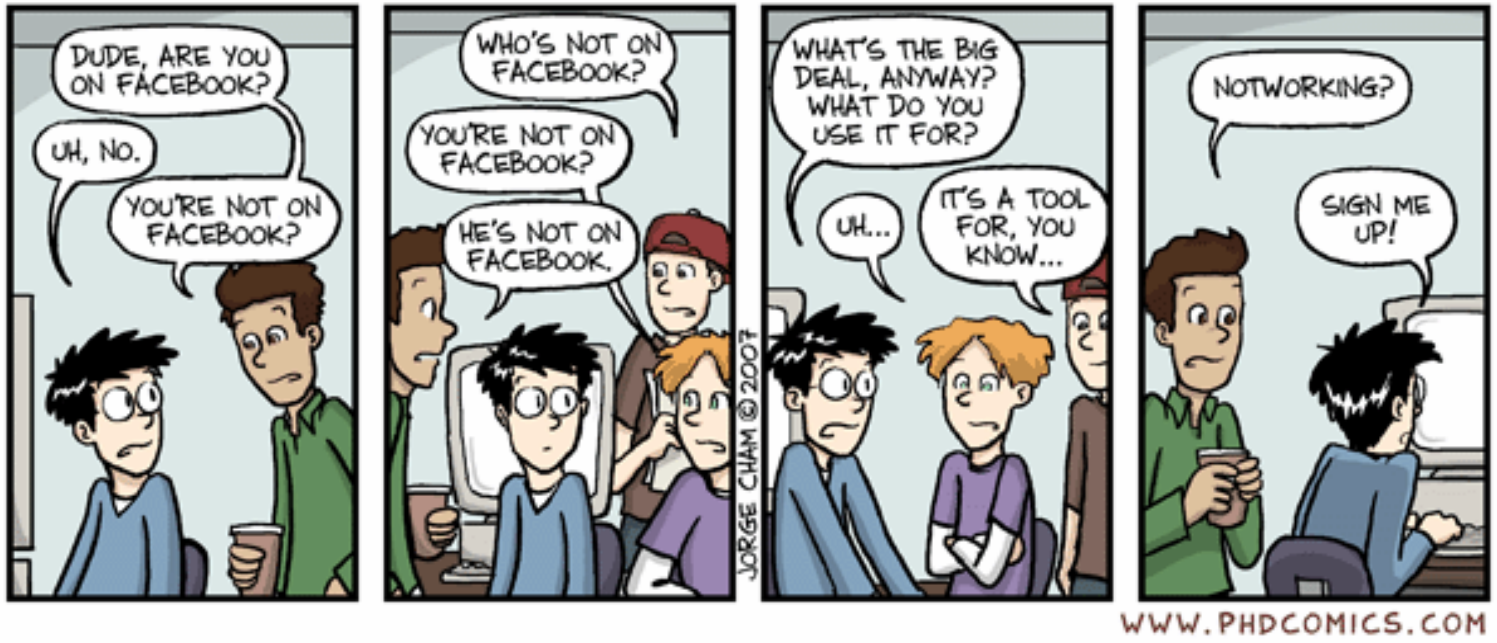
- Structure
  - Aim, self-introduction, lectures, afternoon session
- General session
  - Preliminaries, genetic epidemiological studies, discussion
- Genetic association studies
  - Background, issues, genomewide association, case study
- Some specific topics
  - Retrospective methods, analysis of pathways
- Exercise

Aim

- The first aim is to give a brief overview of genetic analysis of complex traits and diseases in humans
- The focus is on practical issues of data analysis
- It will also include specific examples of genetic association study
- The computer exercises will be both from specialized programs and R
- The expectation is that this will serve as a forum for a range of issues and a contact point for future researches

Self-Introduction

Yourself, your institution and your expectation



We are more positive than this!

The Afternoon Session (by Prof Andrea S Foulkes)

- **LD and HWE.** A general introduction to genetic association studies including linkage disequilibrium (LD) and Hardy Weinberg equilibrium (HWE) are described as well as the R-package genetics and corresponding functions for formally testing for the presence of LD and HWE.
- **Multiple testing adjustments.** These include both single-step and step-down adjustments (e.g. Bonferroni and false discovery rate control) as well as resampling-based methods (e.g. the approaches of Westfall and Young, 1993 and Pollard and van der Laan, 2004) This portion describes applications of existing functions and packages, including p.adjust() and qvalue, as well as alternative, simple coding examples for making appropriate adjustments.
- **Accounting for ambiguity in phase.** Haplotype reconstruction techniques are typically applied to population-level association data, in which allelic phase is generally unobservable. The focus is placed on one such approach that uses an expectation-maximization type algorithm. This approach can be based solely on observed genotype data or additionally incorporate information on a quantitative trait. Here emphasis is placed on the application of functions within the haplo.stats package.
- **High-dimensional data methods.** Two approaches and associated tools for handling the high-dimensional aspect of the data, namely random forests (RFs) and mulitivariate adaptive regression splines (MARS), are described for the genetics setting. Both RFs and MARS are machine learning approaches that represent extensions of the classification and regression tree methodology. Here, applications of functions within the R-packages randomForest and Earth are provided

Preliminaries

- Complex traits
- Genetic epidemiology
- Mathematical statistics
- Statistical computing
- Population genetics

Complex Traits and Approaches to Its Genetic Basis

- Complex traits refer to common diseases or traits with no clear modes of Mendelian inheritance Reduced penetrance, heterogeneity, phenocopy, pleiotrophy,... (Lander & Schork 1994), environmental factors, examples include diabetes, heart diseases, mental disorders, height, body-mass index (BMI)
- Methods include the assessment of familial aggregation for heritability, identification of major gene effect, study of cosegregation of genetic marker with putative disease-predisposing loci in the so-called linkage studies and association studies in search of frequency differences between cases and controls and/or correlation between genotype and phenotype as a quantitative trait. Morton et al. (1983), Khoury et al. (1993), Thomas (2004)

Genetic Epidemiology I

- A hybrid of genetics and epidemiology with evolving framework.
- *The study of the joint action of genes and environmental factors in causing disease in human population and their patterns of inheritance in families* Thomas DC. *Statistical Methods in Genetic Epidemiology.* Oxford University Press 2004
- *In a narrow sense genetic epidemiology is concerned with inherited disease. More broadly, it deals with inherited variation in contemporary populations. Although disease provides a focus, it does not specify the contents which include Mendelian and non-Mendelian mechanisms, normal variation, mapping major and minor genes, map integration, selection, genetic loads, allelic association, population structure, forensic DNA evidence, radiation genetics, and mutation - in short, every aspect of epidemiology and mathematical genetics that is concerned with inheritance and every aspect of population genetics that is not concerned primarily with evolution.* – Morton NE. Human Heredity (2000) 50:5-13
- A science that deals with etiology, distribution, and control of disease in groups of relatives and with inherited causes of disease in populations -- Morton NE (1982) *Outline of Genetic Epidemiology*, Karger, New York.
- Morton et al. (1983). MGE, Khoury et al. (1993), Thomas (2004), Recent reviews on Lancet, Zhao & Tan (2006) Curr Bioinformatics

Genetic Epidemiology II

- Genetics was “genetic epidemiology” as stated in Mendel’s laws, examples include:
  - Fisher’s work on quantitative genetics, Fisher RA (1918) *Trans Roy Soc Edinburgh* 52,399-433, see also the University of Adelaide Digital Library http://digital.library.adelaide.edu.au/coll/special//fisher/index.html
  - Rao’s work on score statistic, Rao CR (2001) JSPI 97:3-7
  - Recent interest in haplotype inference by statisticians and computer scientists
- Summary by Ann Spence (2005) *Encyclopedia of Biostatistics II*, Wiley
  - It is impossible to include all important factors that contributed to the discipline
  - The long-term goal is to understand causes of genetic disease in humans
  - Genetic epidemiology has been responsive to the need for interaction and exchange of ideas, e.g., *Genetic Epidemiology* from 1984, IGES in 1991, GAW (http://www.gaworkshop.org)

Epidemiologic Methods

- Measure of disease (D) –exposure (E) association
  - Relative risk (RR)=P(D|E)/P(D|not E)
  - Odds ratio (OR)=[P(D|E)/P(D|not E)]/[P(D|not E)/P(not D|not E)]
  - Excess risk (ER)=P(D|E)-P(D|not E)
  - Attributable risk (AR)=[P(D)-P(D|not E)]/P(D)=P(E)(RR-1)/[1+P(E)(RR-1)]
- Study designs
  - Population-based studies
  - Exposure-based sampling — cohort studies
  - Disease-based sampling — case-control studies
    - Risk-set sampling, as if stratified or matched by time
    - Case-cohort sampling, OR provides estimate of RR
- Important concepts
  - Bias, confounding, interactions
  - Causal inference
- Statistical modelling
  - Linear and logistic regression and generalized linear models
  - Cox regression

Jewell NP (2004) *Statistics for Epidemiology.* Chapman & Hall

Odds and Odds Ratio

For an event E with probability P(E), the odds of the event is defined as P(E)/(1-P(E)). For two events E_1, E_2, the odds ratio is defined as odds(E_1)/odds(E_2)

		Disease		
		D	not D	
Exposure	E	*a*	*b*	*a+b*
	not E	*c*	*d*	*c+d*
		a+c	*b+d*	*N*

Let D=disease, E=exposure $p_1 = p(D \mid E),\ p_2 = p(D \mid not\, E)$

The Odds Ratio is defined by $OR = \frac{p_1/(1-p_1)}{p_2(1-p_2)} = \frac{p_1(1-p_2)}{p_2(1-p_1)} = \frac{ad}{bc}$

$H_{0:}$:D and E are independent $\Longleftrightarrow$ *OR=1*

We can use contingency table chi-square or logistic regression testing for statistical significance

Mathematical Statistics

- Random variables
- Distribution
- Mean, variance, covariance, correlation
- Significant testing
- Type-I and type-II errors
- Descriptive analysis
- Statistical model
  - Distribution (Normal and Chi-squared)
  - Linear, general linear, generalised linear, mixed models
- Bayes Theorem
- Likelihood
  - Fisher’s information
  - Wald’s, score, and log-likelihood ratio tests
- Bayesian inference

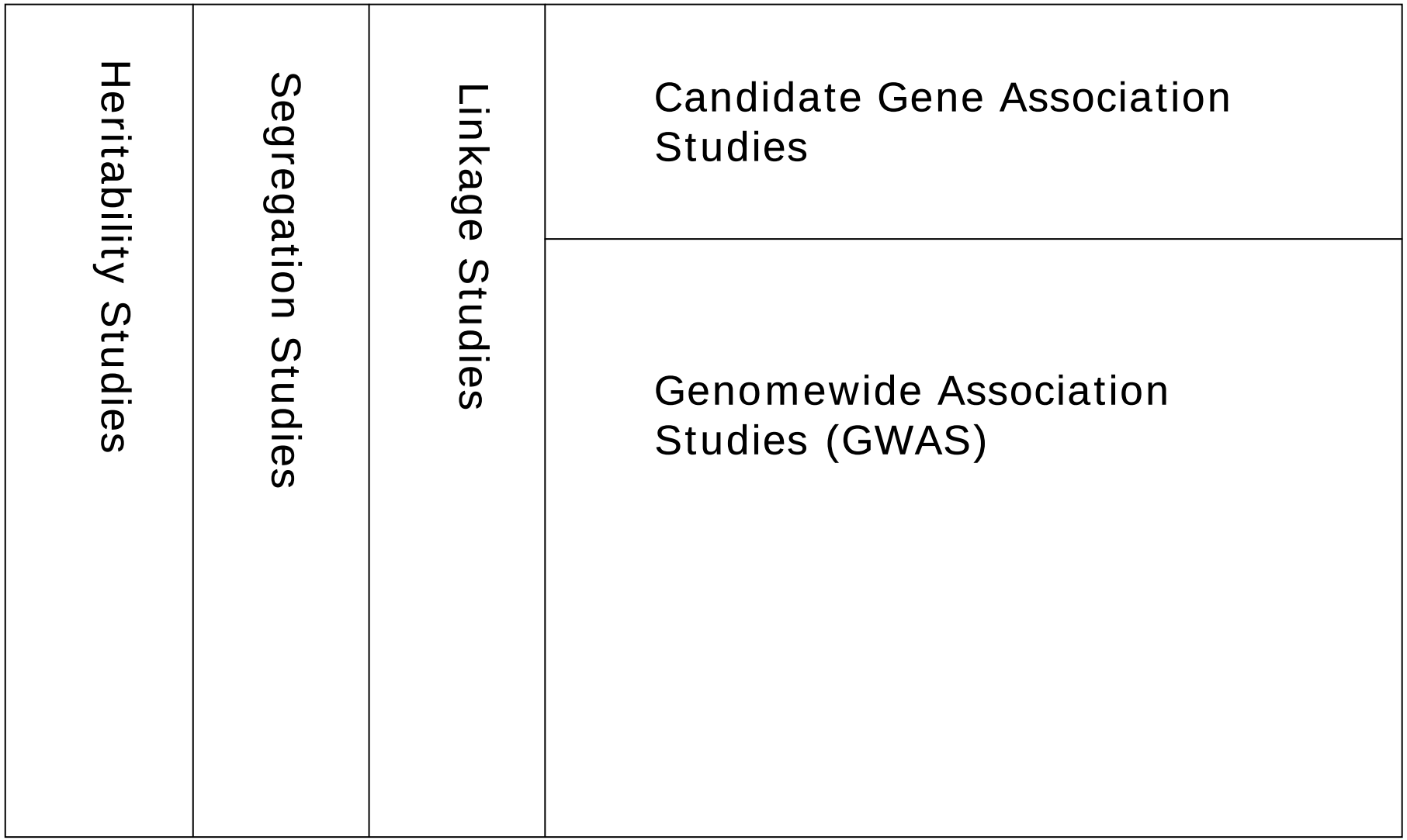
Statistical Computing

- Operating system and utilities
- Networking and Internet
- Standalone applications
- Statistical analysis systems
  - SAS
  - Stata
  - S-PLUS/R
- High-level languages and script programming

Population Genetics

- Chromosomes, DNA, RNA, protein
- Alleles, genotypes, haplotypes
- Hardy-Weinberg equilibrium
- Genetic drift, selection, migration
- Random mating
- Penetrance
- Mendel’s Law
- Linkage disequilibrium and recombination
- Quantitative genetics

Genetic Epidemiological Studies



Heritability Studies

- Genetic heritability is defined for a quantitative trait as the proportion of variation attributable to genetic factors, and extended to categorical traits through reference to a liability model. The value of the genetic heritability varies according to factors taken into account.
  - For a binary trait, such as whether or not an individual has a disease, heritability is not the proportion of disease in the population attributable to or caused by, genetic factors.
  - For a continuous trait, genetic heritability is not a measure of the proportion of an individual's score attributable to genetic factors. Heritability is not about cause *per se*, but about the causes of variation in a trait across a particular population.
- Adoption (rearing of a nonbiological child) studies
- Migrant studies – migrants carry a risk reflecting country of origin
- Twin study – differences between monozygotic and dizygotic twins
- Family studies
- Examples

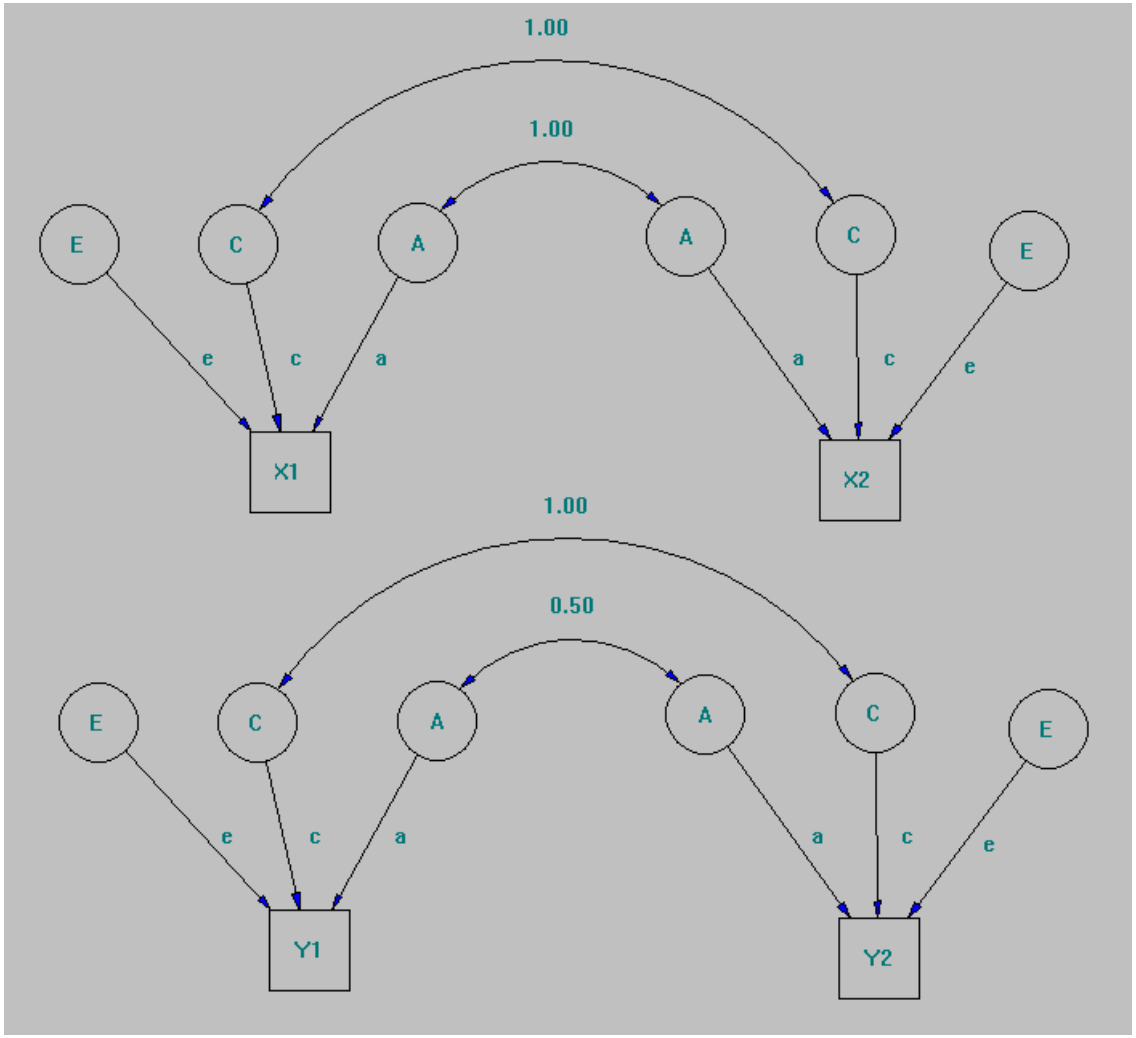
Twin Studies: Path Model

MZ

$r_{MZ} = a^2 + c^2$

DZ

$r_{DZ} = 0.5\,a^2 + c^2$



ACE model

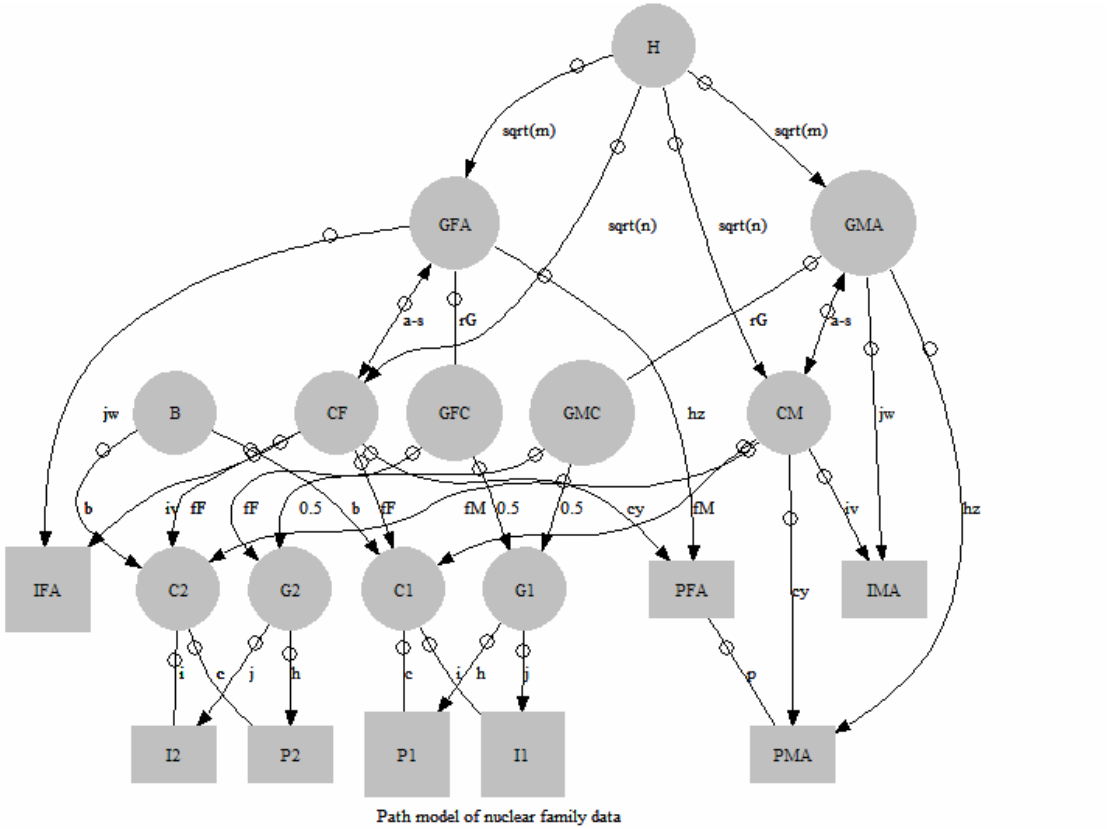
Twin Studies: Heritability Estimate

One can easily derive the covariance or correlation between two types of twin pairs, but an intuitive way is to follow the path-tracing rules. Both approaches give the same answer. Assuming the phenotypes are standardised, this is also the usual heritability estimate. As we can see it requires the assumption of common environment between MZ and DZ twins. Alternatives are available (Feldman et al. 1997, Science) and more assumptions are required (Falconer and Mackay 1996). Its variance of can be obtained, based on that of correlation. The model can be extended to consider gene-environmental interactions (Guo 2000, Hum Hered).

The heritability and variance are $a^2 = 2(r_{MZ} - r_{DZ})$

$$4\left[\left(1-r_{MZ}^2\right)^2 / n_{MZ} + \left(1-r_{DZ}^2\right)^2 / n_{DZ}\right]$$

Path analysis of Nuclear Family Data



Path model of nuclear family data

Transmission of mixed homogamy in nuclear families. P, G, C, I indicate phenotype, genotype, indexed environment and index, respectively. H=homogamy (mating), B=non-transmitted common sibship environment, F=father, M=mother, 1=child 1, 2=child 2, whereas h, c, j, b indicate path coefficients in children and z, y, w, x are the adult version. Morton et al. 1983, p28

Software

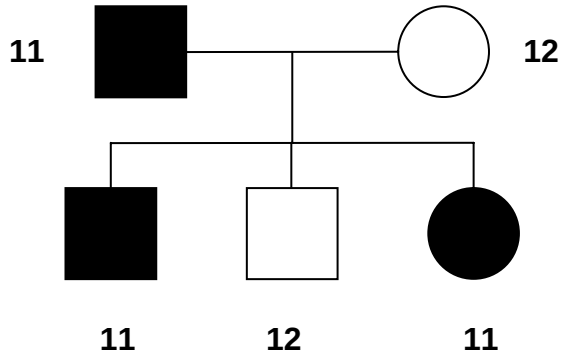
- FISHER
- PATHMIX
- LISREL, Mx, Mplus
- Recast in gllamm framework so that they can be modelled through Stata and S-PLUS/R (Rabe-Hesketh et al. (2008) Biometrics 64:280-8)

Segregation Studies: Overview

- The inference of mode of inheritance through analysis of family data. It is most commonly conducted in likelihood framework. Complex segregation analysis considers multiple mating types and multiple determinants.
  - Monogenic if a trait is dominated by effects of a single locus
  - Oligogenic if a trait is influenced by a few loci
  - Polygenic if a trait is influenced by a large number of loci with small effects
- It involves three types of parameters: Allele frequencies, Penetrances (genotypic or ousiotypic means), Transmission probabilities
- Advantage and disadvantages: Best-fitting model is not necessarily correct AND inclusion of genetic markers increases the power of genetic analysis, Nevertheless it is fundamental to the understanding of linkage analysis
- Morton et al. (1983) Methods in Genetic Epidemiology
- Khoury et al. (1993) Fundamentals of Genetic Epidemiology
- Thomas (2004) Statistical Methods in Genetic Epidemiology

Segregation Analysis: A Simple Example

Consider a nuclear family



- Let
  - p=allele frequencies,
  - f=penetrances,
  - T=transmission probabilities given parental genotypes
- The likelihood can be written as follows
- L = p(11)f(11) x p(12)(1-f(12)) x T(11|11,12)f(11) x T(12|11,12)(1-f(12)) x T(11|11,12)f(11)

Applications to Breast Cancer

- Claus et al. (1991) showed a dominant major gene model is more preferable to pure environmental, pure polygenic, or recessive major gene models
- Andrieu and Damenais (1997) using Bonney's class D model showed women with a young age at menarche have dramatically higher penetrances than those with older ages at menarche
- This led to the finding of BRCA1 and BRCA2 genes
- A breast and ovarian analysis of disease incidence and carrier estimation algorithm (BOADICEA) is available http://www.srl.cam.ac.uk/genepi/boadicea/boadicea_home.html as with R/Bayesmendel, http://www.cancerbiostats.onc.jhmi.edu/BayesMendel

More details from Thomas (2004) *Statistical Methods in Genetic Epidemiology* and Blangero J (2005) in Encyclopedia of Biostatistics II.

Software

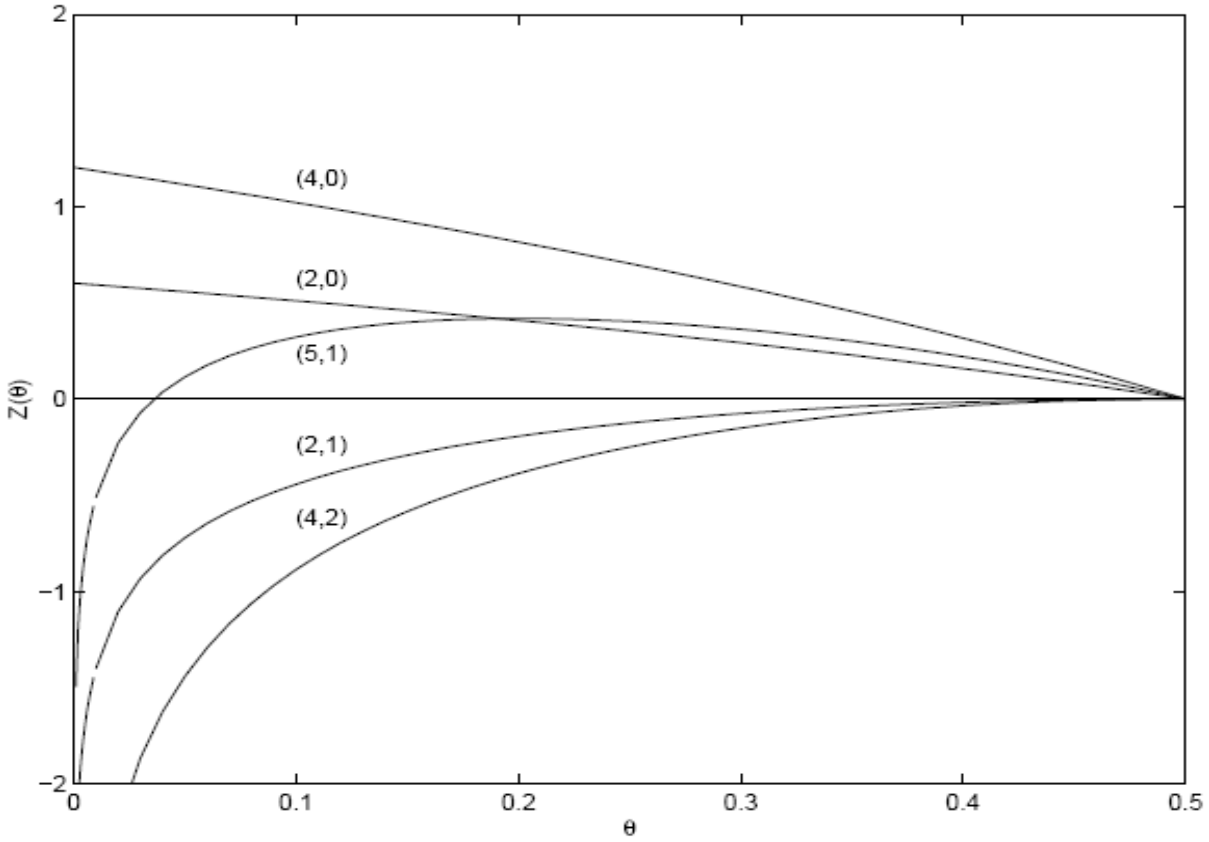
- POINTER
- MENDEL
- PAP
- SAGE
- SOLAR

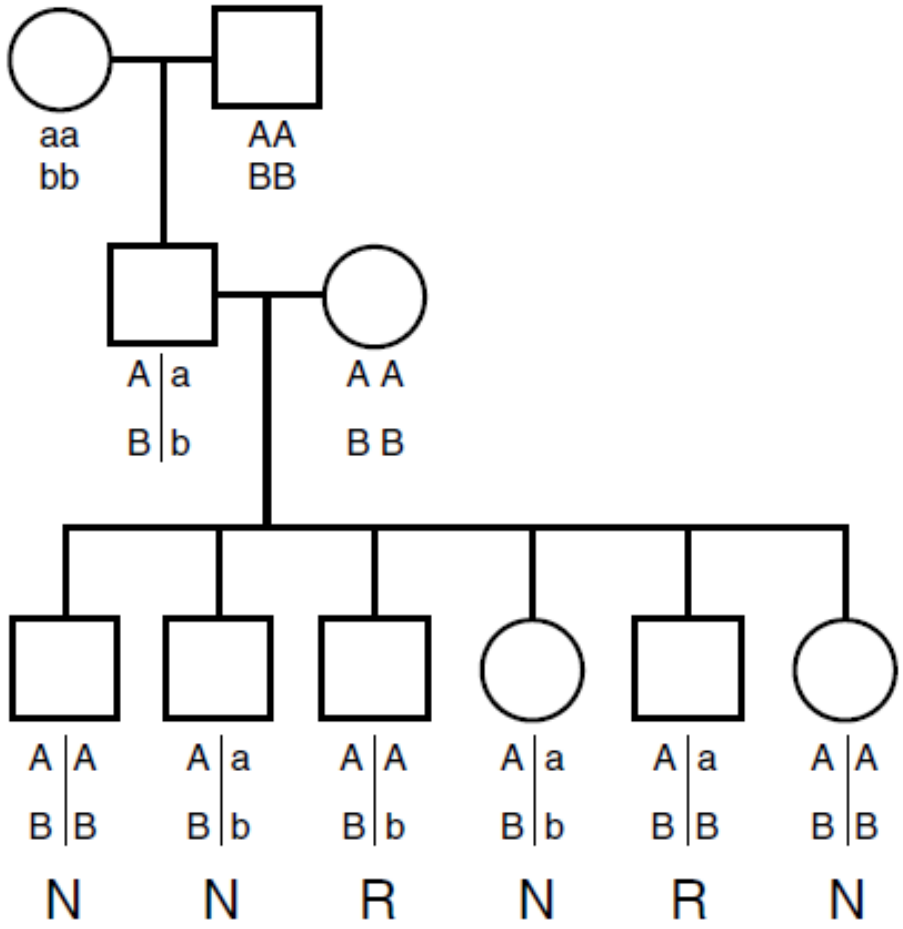
Linkage Studies

- Aim and scope
  - To study frequency of cosegregation of trait and alleles at putative marker loci in families
  - It is desirable since for disease with low penetrance, it is more likely familial aggregation would be more genetic rather than environmental; data are likely to be more accessible and thus have played a key role in positional cloning
- The statistical model is encapsulated in segregation analysis but with marker information, map function is also relevant. Moreover, methods for detecting marker-disease association through linkage disequilibrium are special cases of model-based linkage analysis.
- Model-based method refers to the situation when the mode of inheritance (allele frequency, penetrance) is fully specified, otherwise it is model-free

Direct Counting

- If all meioses *(n)* in a pedigree can be classified as recombinants *(r)* or nonrecombinants *(n-r)*, the likelihood function is simply the binomial probability function, $L(\theta) = \binom{n}{r} \theta^r (1-\theta)^{n-r}$ and LRT $\Lambda(\theta) = 2L(\theta)/L(0.5)$
- The lod score is defined as $Z(\theta) = \log 10[L(\theta)/L(0.5)]$, so that a value of 3 means the observed data is 1000 times more likely than that under the null.
- The lod score curves can be drawn based on this.



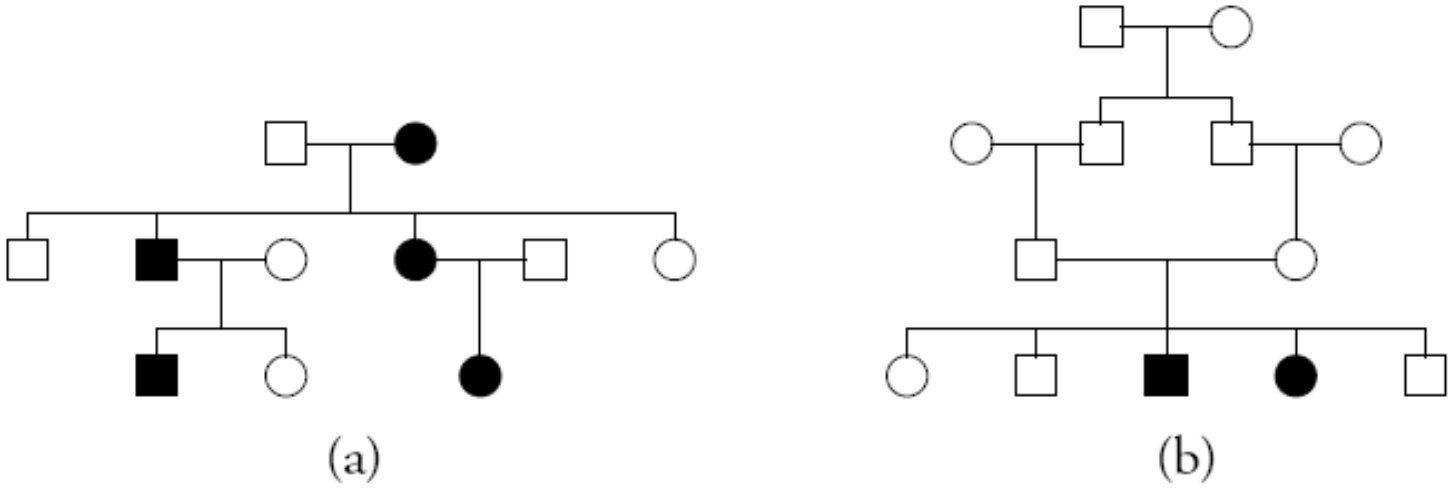


$$Z(\theta) = \log_{10} \frac{\theta^2(1-\theta)^4}{(0.5)^2(0.5)^4} = \log_{10} 2^6\theta^2(1-\theta)^4$$

A phase-known pedigree with two codominant loci with alleles A/a and B/b. Nonrecombinant and recombinant meioses are marked with N and R, respectively. The MLE of recombination rate (θ) is 2/6 and lod score 0.1475

For phase-unknown pedigree(s), one need to consider all possible phases and parameter estimation is customarily done through maximum likelihood

Disease Models



Two pedigrees with typical (a) autosomal dominant inheritance and (b) autosomal recessive inheritance. Affected individuals have filled symbols

Parametric Linkage Analysis Using Extended Pedigrees

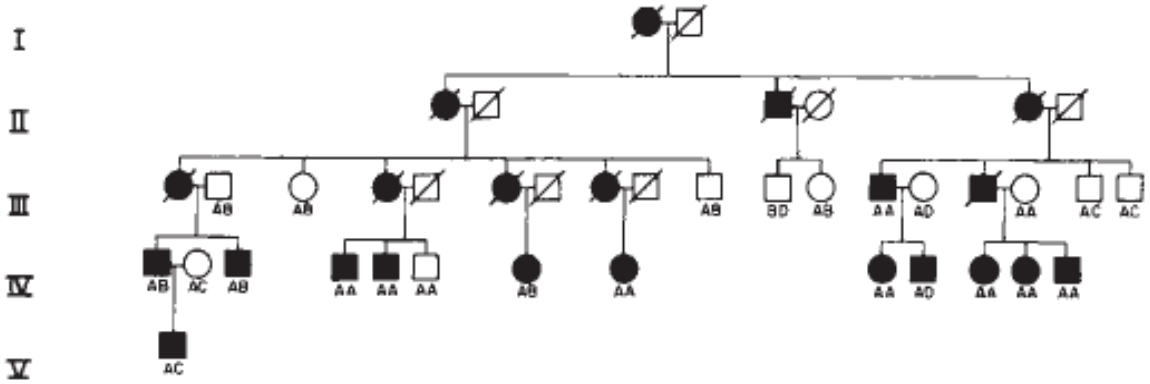
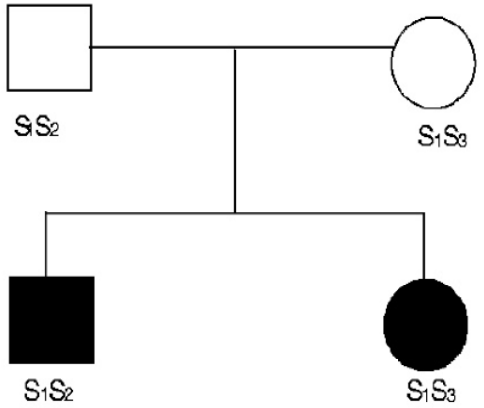


Fig. 1 Pedigree of an American Huntington's disease family. Symbols: circles, females; squares, males; a black symbol indicates that an individual is affected with Huntington's disease; a slashed symbol indicates that an individual is deceased. This pedigree was identified through the National Research Roster for Huntington's Disease Patients and Families at Indiana University. Relevant family members were examined by a neurologist and blood samples were obtained. EBV-transformed lymphoblastoid cell lines were established for the individuals whose genotypes are shown and have been stored at the Human Genetic Mutant Cell Repository, Camden, New Jersey. Phenotypes at the G8 locus shown under each symbol were determined by Southern blotting as outlined in Fig. 3. For the purposes of confidentiality, selected individuals are not shown.

Gusella et al. (1983). Nature

Affected Sib-pair (ASP) Methods



$$P(IBD = 0) = \frac{1}{4} - \frac{(\Psi - 0.5)V_A + (2\Psi - \Psi^2 - 0.75)V_D}{4(K^2 + 0.5V_A + 0.25V_D)}$$

$$P(IBD = 1) = \frac{1}{2} - \frac{2(\Psi^2 - \Psi + 0.25)V_D}{4(K^2 + 0.5V_A + 0.25V_D)}$$

$$P(IBD = 2) = \frac{1}{4} + \frac{(\Psi - 0.5)V_A + (\Psi^2 - 0.25)V_D}{4(K^2 + 0.5V_A + 0.25V_D)}$$

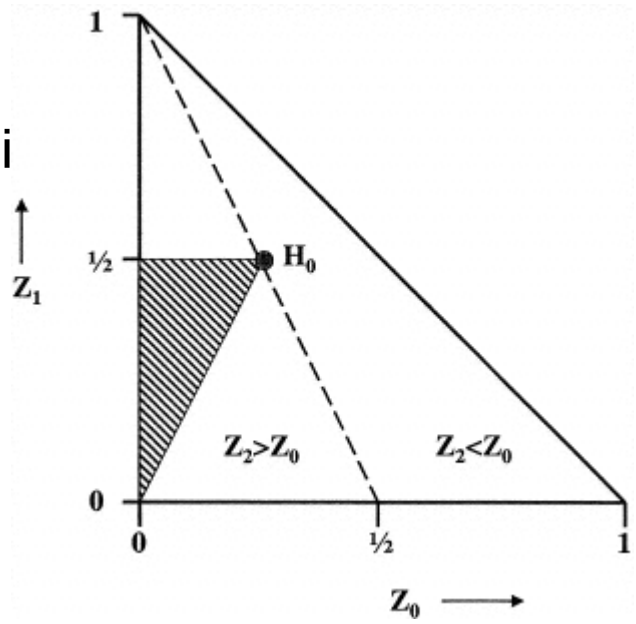
Tests using ASPs

- Simple χ^2 test $S_1 = \sum_i (o_i - e_i)^2 / e_i$

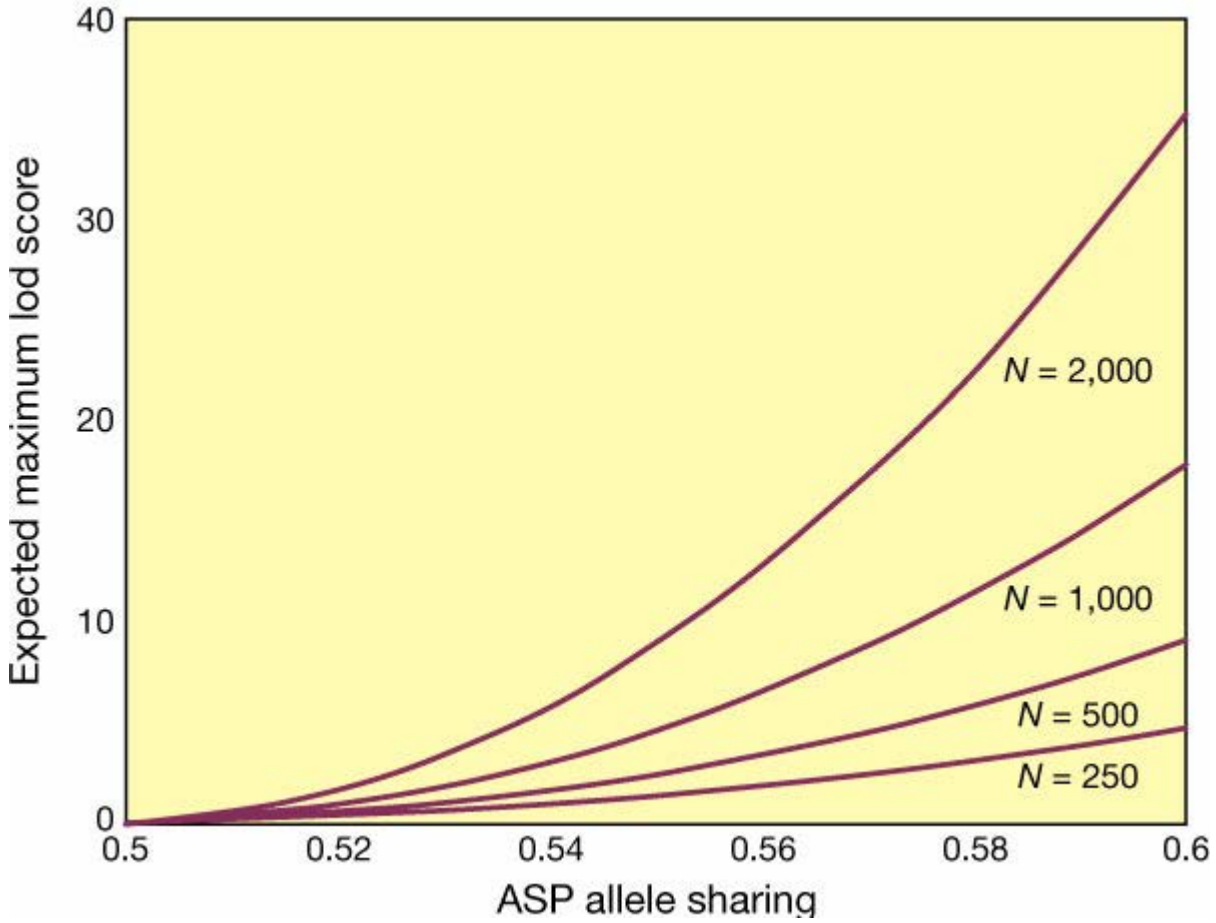
- Means test $S_2 = \left(\frac{1}{2} n_1 + n_2 \right) \Big/ \sqrt{\frac{n}{8}} \sim N(0,1)$

- MLS $L = \sum_i z_i P(x \mid z_i)$

Sham & Zhao (1998) in Bishop MJ (Ed) Gui To Human Genome Computing, 2nd Edition



Elston (2001) AJHG 69: 1149-50



Range of number of ASPs required to detect linkage as a function of allele sharing. Risch (2000) Nature

Three Affected Siblings

Possible Genotype Configurations (n=20):

1/3	1/3	1/3
1/3	1/3	1/4
1/3	1/3	2/3
1/3	1/3	2/4
1/3	1/4	1/4
1/3	1/4	2/3
1/3	1/4	2/4
1/3	2/3	2/3
1/3	2/3	2/4
1/3	2/4	2/4
1/4	1/4	1/4
1/4	1/4	2/3
1/4	1/4	2/4
1/4	2/3	2/3
1/4	2/3	2/4
1/4	2/4	2/4
2/3	2/3	2/3
2/3	2/3	2/4
2/3	2/4	2/4
2/4	2/4	2/4

Inheritance Vectors (n=64):

1: 000000
2: 000001
010000
000100
3: 000010
100000
001000
4: 000011
001100
110000
5: 000101
010100
010001
6: 000110
011000
010010
001001
100100
100001
7: 000111
011100
010011
001101
110100
110001
8: 001010
101000
100010
9: 001011
101100
100011
001110
111000
110010
10: 001111
111100
110011
11: 010101
12: 010110
011001
100101
13: 010111
011101
110101
14: 011010
101001
100110
15: 011011
101101
110110
011110
100111
111001
16: 011111
111101
110111
17: 101010
18: 101011
101110
111010
19: 101111
111110
111011
20: 111111

Equivalent Inheritance Vectors:

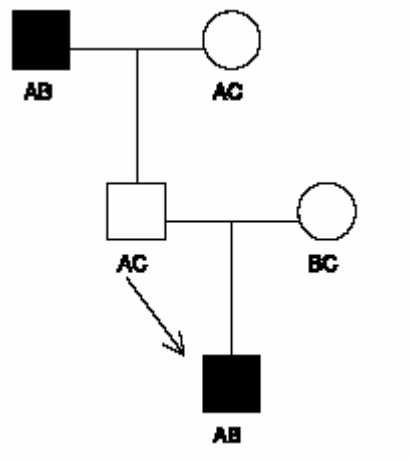
1,11,17,20
2,3,5,8,13,16,18,19
4,10,12,14
6,7,9,15

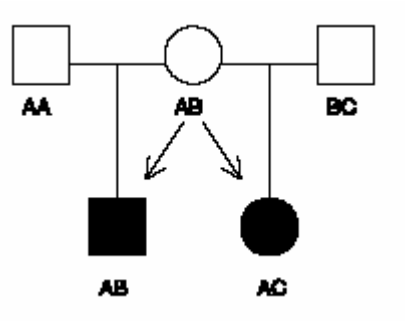
Pair IBD Sharing:

IBD=2,2,2	4/64
IBD=2,1,1	24/64
IBD=2,0,0	12/64
IBD=1,1,0	24/64

Nyholt (2000) Am J Hum Genet

Affected Relative Pairs



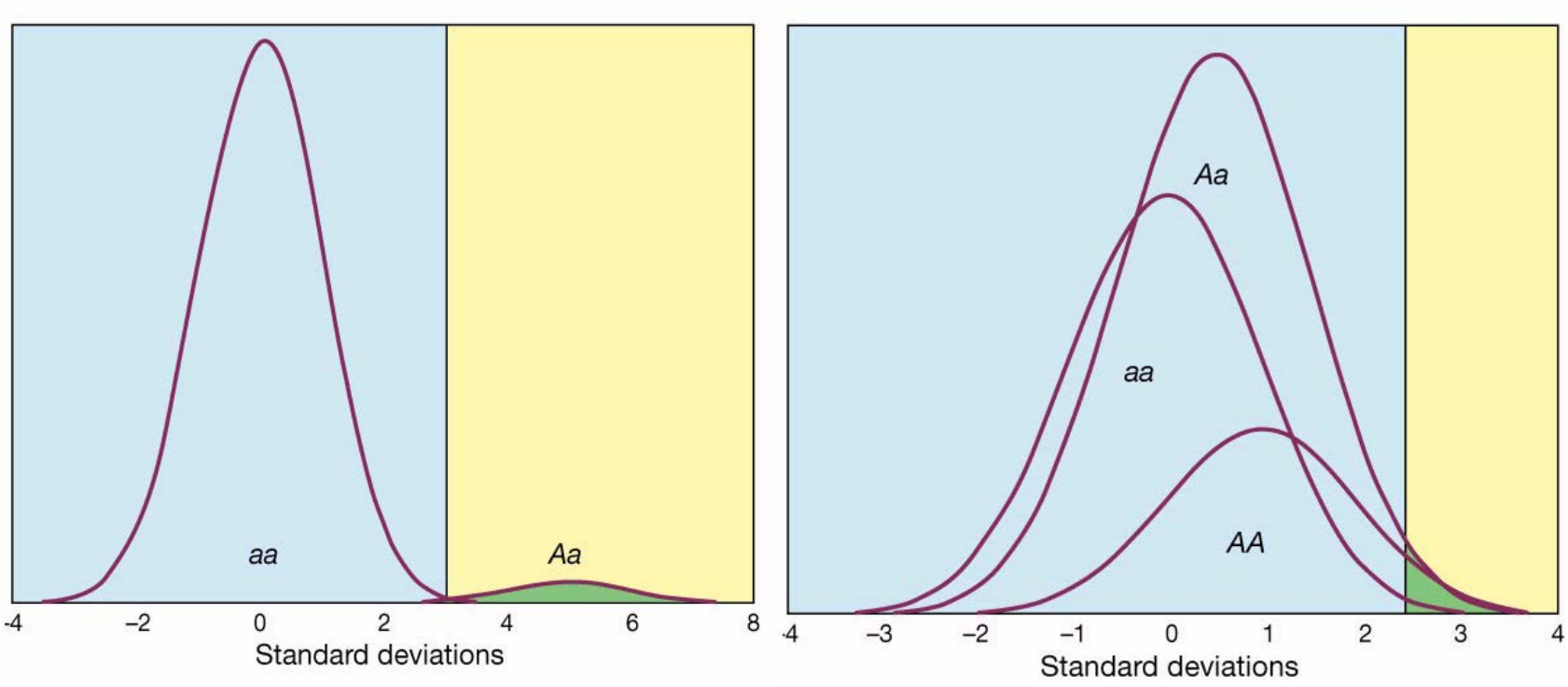


IBD if sharing the same allele of a ancestor
IBS if sharing the same allele, regardless the ancestral origin

Quantitative Traits

Mendelian (dominant)

Non-Mendelian (additive)



Risch (2000) Nature

Quantitative Trait Loci

- Haseman-Elston regression for sibling data
- Variance component model
- Regression conditional on trait
- Generalised estimating equations (GEE)

Further Issues in Linkage Studies

- Linkage heterogeneity
- Imprinting
- Two-locus model
- Genomewide significance, the Lander-Botstein model
- Disadvantages
- Design of linkage studies
- Staged design

Software

- LIPED, LINKAGE, FASTLINK, MFLINK, VITESSE
- SLINK/FASTSLINK
- MENDEL, SIMLINK
- SUPERLINK
- MAPMAKER/GENEHUNTER, Allegro, MERLIN
- ASPEX, SPLINK
- MORGAN

Evolution of Computer Implementations

- Algorithms evolution span over ~50 years
- Standalone software programs (Pascal, Fortran, C/C++,...) (http://inkage.rockefeller.edu)
- Increasingly limited and requires marriage with established systems (commercial and free, e.g. SAS/Stata/S-PLUS and R).
- In particular for R
  - The call for integration is reminiscent of the development of Linux system
  - R is a general computing environment for both practitioners and programmers. Standard statistical and genetic analyses can be conducted in a reliable and unified fashion. It makes efficient statistical modelling possible.
  - It is accessible and free
  - It is easy to extend and a large number of packages maintained by a large and dedicated team
  - Zhao & Tan. *Hum. Genomics*, **2**, 258-265, 2006

SAS

- A systems (http://www.sas.com, http://en.wikipedia.org/wiki/SAS) for many fields of research
- Database support such as Oracle, MySQL and facility such as ODBC, data visualisation, fantastic data subsetting facility e.g.,

```
PROC SQL;
    CREATE TABLE mytable AS SELECT * FROM  master
    WHERE rsn IN (SELECT rsn FROM d1) & id IN (SELECT id
    from d2);
QUIT;
```

- Modelling facilities, statistics, operations research
- SAS/Genetics
  - which includes procedures ALLELE, CASECONTROL, FAMILY, HAPLOTYPE, HTSNP, INBREED, PSMOOTH. General routines in SAS such as MULTTEST and LOGISTIC are better known.

Stata

- Information available from http://www.stata.com
- The most popular software for epidemiologists, and researchers in other fields such as econometrics. Its facilities range from basic data management, computer graphics, programming, to methods for complex survey, longitudinal data analysis and multilevel models. Beside its simple but flexible syntax and comprehensive on-line documentation, it has many unique features such as frequency, probability and analytic weights.
- Support for databases, Internet, multiprocessor
- Stata Journal, Bulletin, a large repository and user community

R

- Linkage analysis
  - Power (powerpkg, gap)
  - Check of relationships (Relcheck)
  - Pedigree analysis (identity, MasterBayes,multic)
  - Plot a genome scan (lodplot)
  - Survival models (kinship)
  - Meta-analysis (Leiden)
- Generic packages as given in ctv for genetics
  - genetics (Warnes. *R News*, **3**, 9-13, 2003)
  - haplo.stats (Lake et al. *Hum. Hered.*, **55**, 56-65, 2003)
  - gap (Zhao J. Stat Soft)
  - Others, LDheatmap, luca, dgc.genetics
  - Multiple testing: qvalue (multtest, twilight, locfdr,...)
  - Bioinformatics: R News 6/5 (http://cran.r-project.org/doc/Rnews/Rnews_2006-5.pdf)
- Packages for genomewide association studies (GWAS)
  - SNPassoc (Gonzalez et al. *Bioinformatics*, **23**, 644-645, 2007)
  - GenABEL (Alchenko et al. *Bioinformatics*, **23**, 1294-1296, 2007)
  - SNPmatrix (Clayton and Leung, *Hum. Hered.*, **64**, 45-51, 2007)
  - pbat2

Discussion

- What are the paradigms in your area?
- Questions?
- Comments?

Genetic Association Studies

- A review of paradigm
- Types
  - Candidate gene vs hypothesis-free approaches
  - Family-based vs population-based
- Model
- Advantage and disadvantage
- Software
- Example

Steps in Positional Cloning

Lander & Schork. Science 265:2037-2048,1994

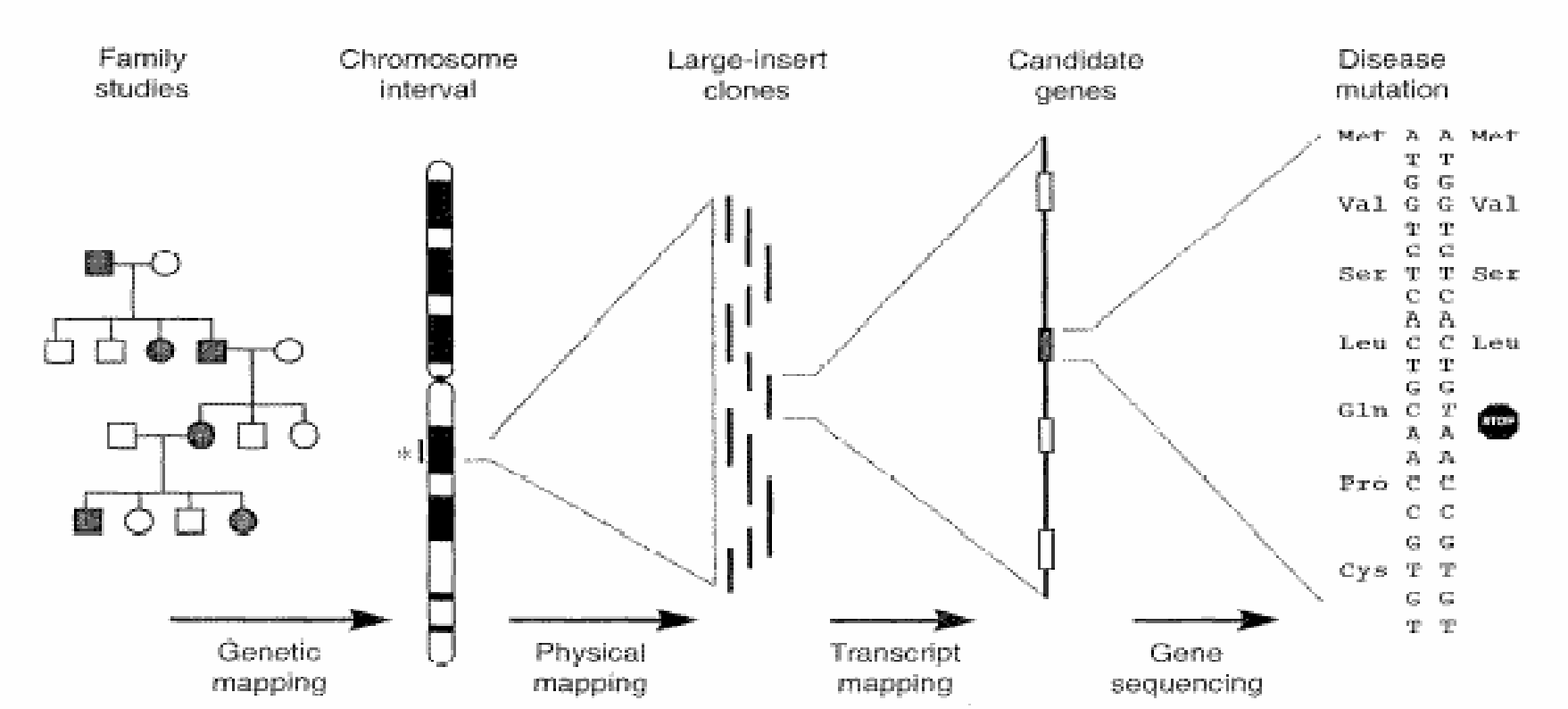


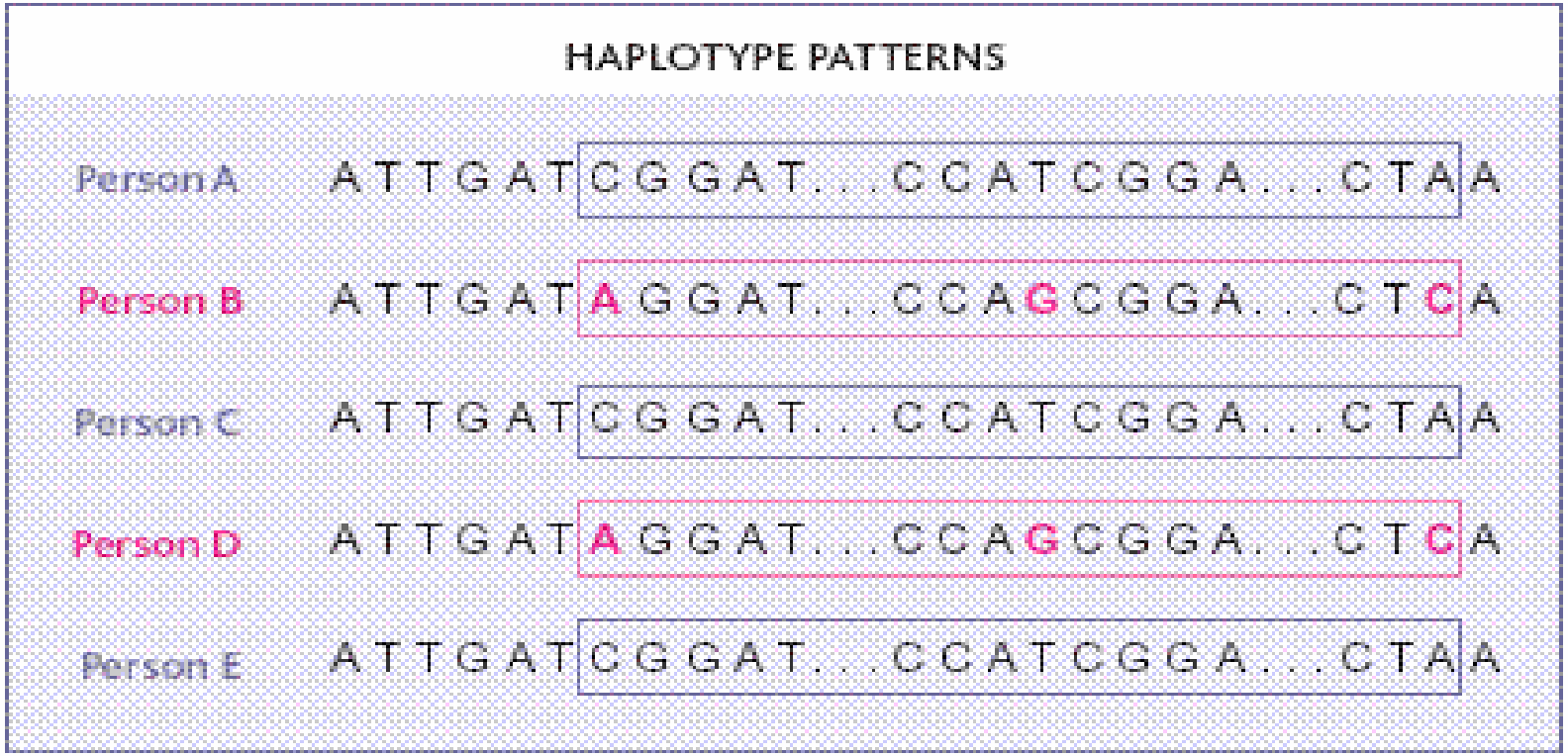
Fig. 1. Steps in positional cloning. Positioning of disease loci to chromosomal regions with genetic markers has become increasingly straightforward, particularly given the recent release of the Généthon genetic map containing 5264 markers (*17*). However, identification and evaluation of the genes within the implicated region remains a major stumbling block.

Schuler GD, et al.(1996) Science 274:540-546, 1996

The Change of Paradigms

- Traditional (e.g. segregation and linkage) methods in human gene-trait study gradually move towards association study.
- It is hugely important but costly
- Large-scale studies are required to maintain statistical power, and it is also necessary to maintain a good coverage of the genome
- Both diseases and traits are under scrutiny
  - The Wellcome Trust Case-Control Consortium (WTCCC), http://www.wtccc.org.uk
  - The Genetic Investigation of Anthropometric Traits (GIANT) consortium (height, body-mass index (BMI, kg/m^2))

Haplotype-trait association

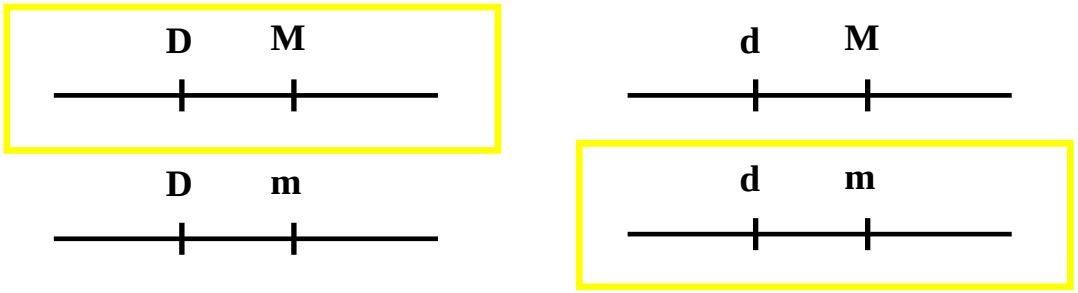


Building blocks. Persons B and D share a haplotype unlike the other three, characterized by three different SNPs.

Couzin (2002) *Science*

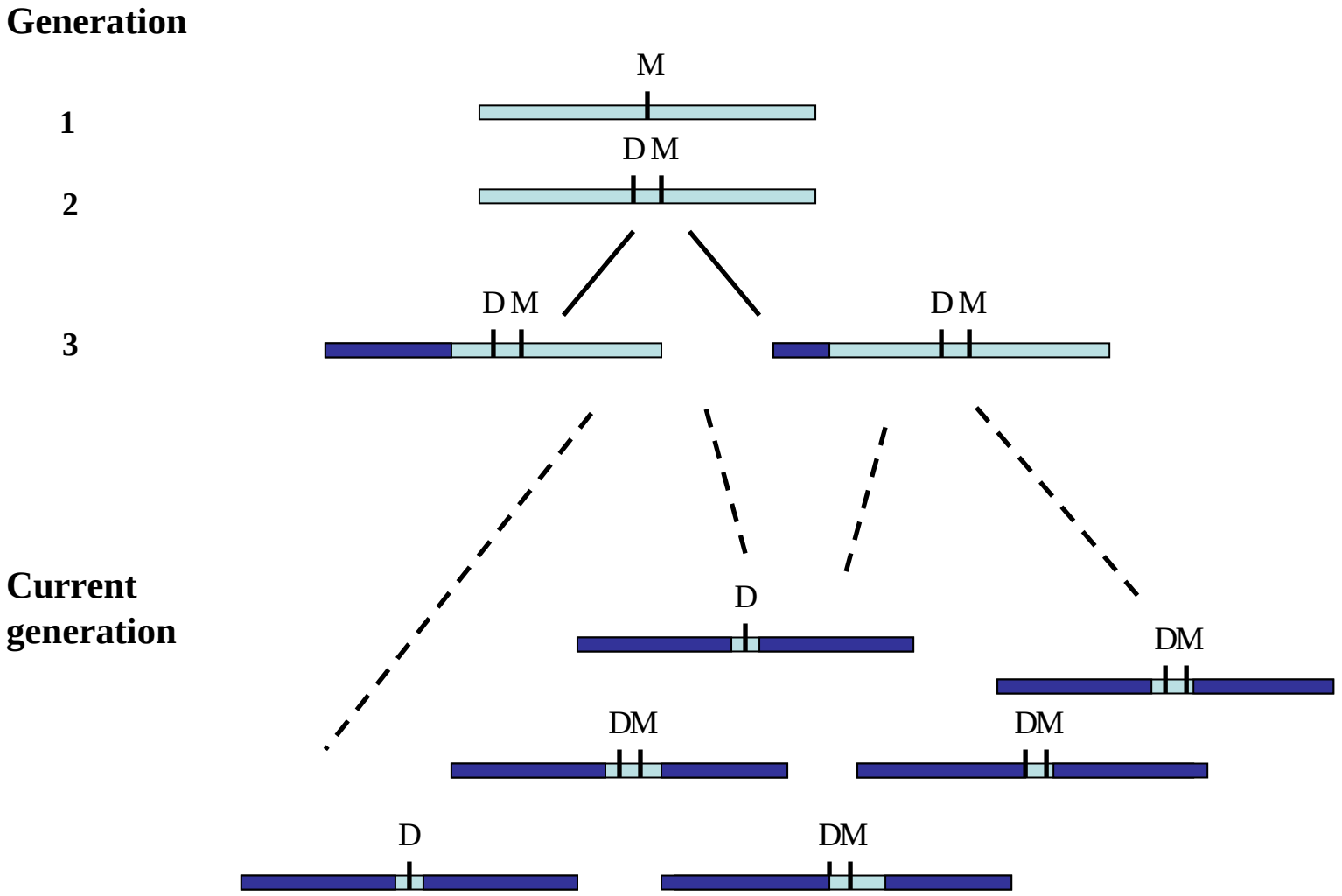
Testing for association with SNPs

Haplotypes



- Power to detect association with marker M depends on
  - frequency of D, P(D)
  - frequency of M, P(M)
  - LD between D and M, P(M|D)
  - risk of disease for genotypes Dd, DD, γ, γ^2

Linkage disequilibrium



Measuring LD

- The expectation of $D = 0$

$$D_{AB} = f_{AB} - f_A f_B$$

- r^2 correlation coefficient
  - Range [0,1]
  - Hill & Robertson (1968)

$$r^2_{AB} = \frac{D^2_{AB}}{f_A f_a f_B f_a} = \rho^2_{AB}$$

- D'
  - Range [0,1]
  - Lewontin (1964)

$$|D'_{AB}| = \text{if } (D > 0)\ \frac{D_{AB}}{\min(f_A f_b, f_a f_B)}$$

$$\text{else } \frac{-D_{AB}}{\min(f_A f_B, f_a f_b)}$$

- Odds-ratio formulation
  - Devlin & Risch (1995)

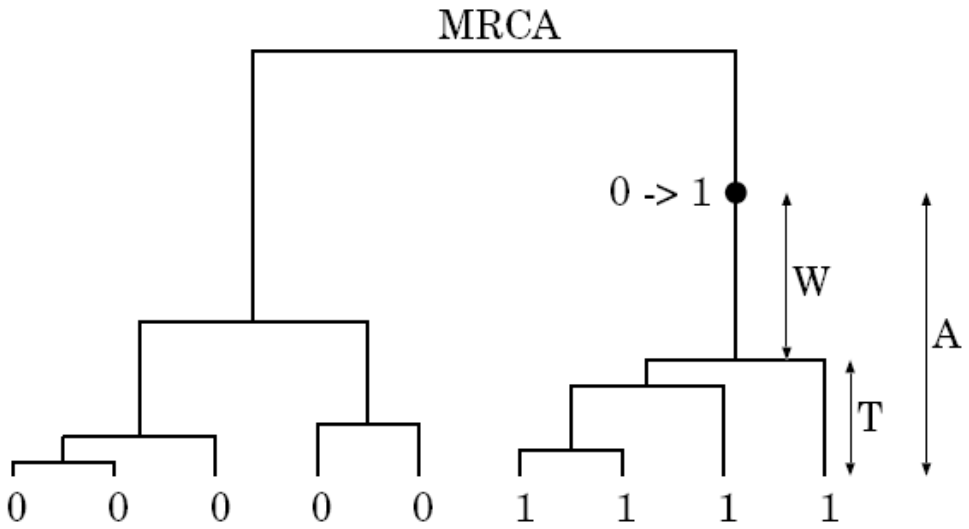
$$\delta_{AB} = \frac{D_{AB}}{f_B f_{ab}}, D_{AB} > 0$$

- For the Malecot model

$$E(D') = Ae^{-td} + B, (1-\theta) \approx e^{-d}$$

Variances of r and D' available from 2LD and LD22 in R/gap

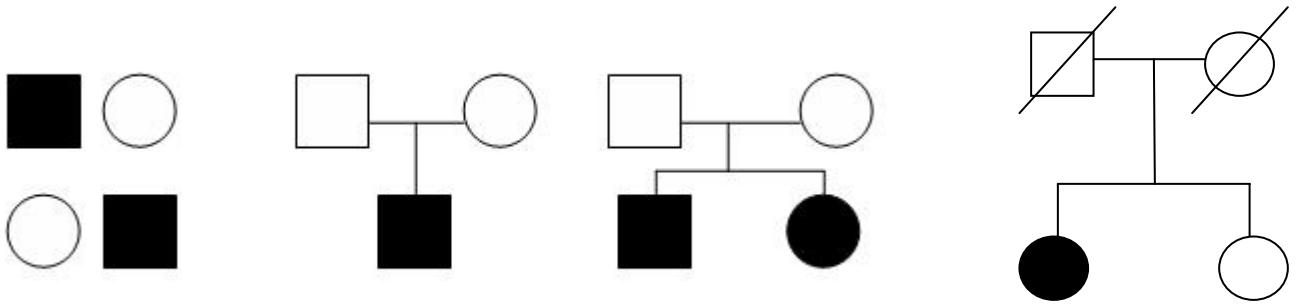
Coalescent Models



Allelic genealogy where a single mutation (solid circle) $0 \longrightarrow 1$ gives rise to a sample of 4 chromosomes carrying the new 1 allele. Indicated are MRCA of the whole sample at time T_{MRCA}, the MRCA of all chromosomes with the 1 allele at time T, the age of the mutation A and the difference between the allelic MRCA and the allelic age W.

Study Designs

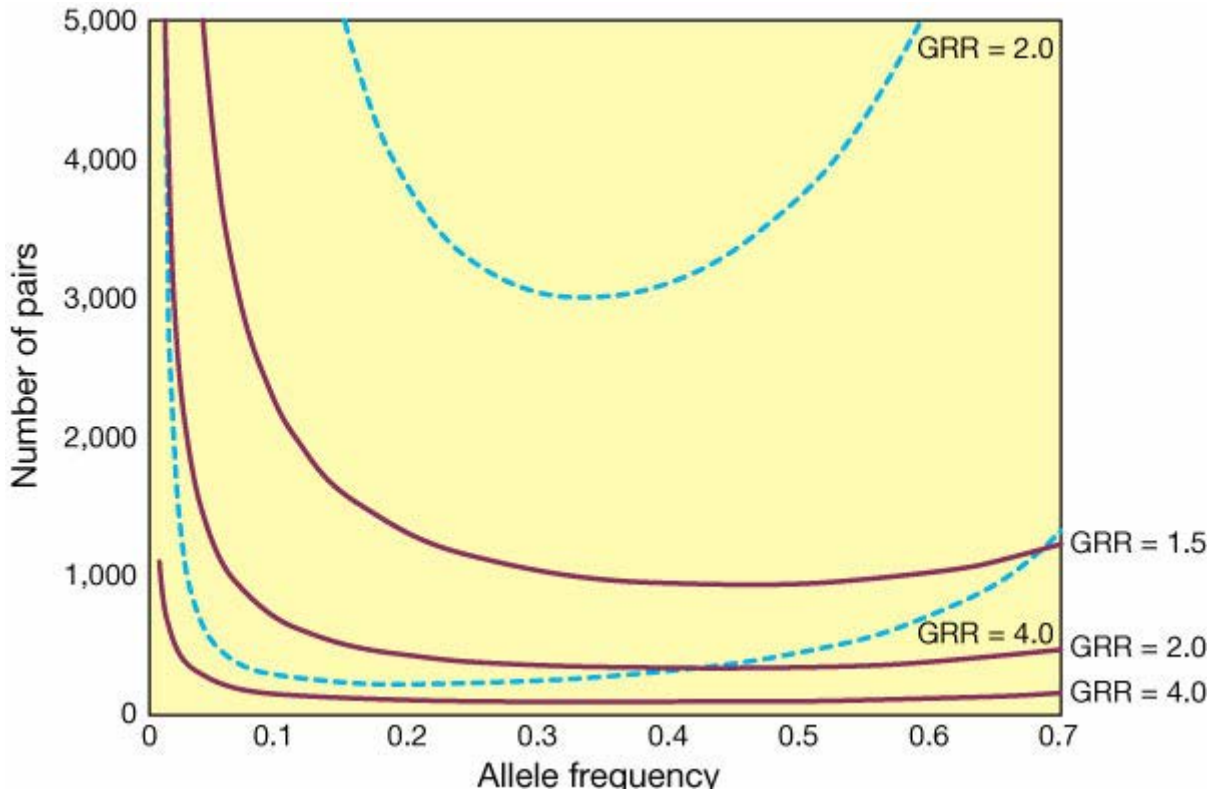
- The impact of technological advances
  - RFLP, SSR, SNP
- Types of samples
  - Three common genetic association designs involving unrelated individuals (left), nuclear families with affected singletons (middle) and affected sib-pairs (right). When parental information is unavailable, an unaffected sibling is used. Males and females are denoted by squares and circles with affected individuals filled with black and unaffected individuals being en



- Statistical considerations: type of population, type-I error, type-II error, disease models (allele frequencies and penetrances), test statistics

		Linkage			Association				
γ	p	Y	N_{asp}	P_A	Het_1	N_{tdt}	Het_2	$N_{asp/tdt}$	$N_{asp}(\times)$
4.0	0.01	0.520	6400	0.800	0.048	1098	0.112	235	4260
	0.10	0.597	276	0.800	0.346	150	0.537	48	185
	0.50	0.576	445	0.800	0.500	103	0.424	61	297
	0.80	0.529	3022	0.800	0.235	222	0.163	161	2013
2.0	0.01	0.502	445835	0.667	0.029	5823	0.043	1970	296710
	0.10	0.518	8085	0.667	0.245	695	0.323	264	5382
	0.50	0.526	3751	0.667	0.500	340	0.474	180	2498
	0.80	0.512	17904	0.667	0.267	640	0.217	394	11917
1.5	0.01	0.501	6943229	0.600	0.025	19320	0.031	7776	4620807
	0.10	0.505	101898	0.600	0.214	2218	0.253	941	67816
	0.50	0.510	27040	0.600	0.500	949	0.490	484	17997
	0.80	0.505	101898	0.600	0.286	1663	0.253	941	67816
Alzheimer's:									
4.5	0.15	0.626	163	0.818	0.460	100	0.621	36	109

Y=probability of allele sharing, P_A=Allele transmission probability.
Power comparison of affected sib-pair linkage and association according to Risch & Merikangas (1996) Science, Zhao (2007) J Stat Soft



Linkage is based on ASPs with a completely linked and informative marker. Association is based on case-control pairs testing the causative locus. Results are obtained via multiplicative models with different genotypic relative risks (GRRs). Risch (2000) Nature

Case-control designs

		K			
γ	p	1%	5%	10%	20%
4.0	0.01	46638	8951	4240	1885
	0.10	8173	1569	743	331
	0.50	10881	2089	990	440
	0.80	31444	6035	2859	1271
2.0	0.01	403594	77458	36691	16307
	0.10	52660	10107	4788	2128
	0.50	35252	6766	3205	1425
	0.80	79317	15223	7211	3205
1.5	0.01	1598430	306770	145312	64583
	0.10	191926	36835	17448	7755
	0.50	97922	18793	8902	3957
	0.80	191926	36835	17448	7755

Table 5: Estimated sample sizes required for association detection using population data.

Multistage Design

- fs=fraction of sample at stage one, fm=fraction of markers at stage two
- Satagopan (2002) fs=25%, fm=10%
- Satagopan (2004) fs=50%, fm=10%, fixed sample size and Armitage trend test
- Satagopan & Elston (2003), fm=15~25%, for a given power
- Skol et al. (2006) fs=10~50%, fm=1~10%, joint analysis is better
- Wang et al. (2006) Affy500k fs=30-40%,fm=0.4-0.5%
- Elston et al. (2007) Ann Rev Genomics Hum Genet

Analytical Issues

- Basic analysis
- Population stratification
  - Genomic controls
  - Structured association
- General issues
  - Balding. *Nat. Rev. Genet*, **7**, 781-791, 2006
  - Elston & Anne Spence. *Stat Med.*, **25**, 3049-3080, 2006
- Specific issues
  - Haplotype-tagging and imputation
  - Multiple-testing, but unlike gene-expression data to use q-value/pFDR
  - Population stratification and outlier detection
  - Genotyping error, calling algorithms

Basic Analysis

- Genotypewise versus allelewise analysis
- Armitage trend test
- Relationship to logistic regression
- Bayesian statistics

Single-locus Cochran-Armitage Trend Tests

- Assuming the sample to be typed at a SNP marker of interest, we can represent genotype data in a 2 x 3 contingency table.
- The Cochran-Armitage trend test of association between disease and the marker SNP is given by

$$X^2 = \frac{\left[\left(p_{2A} + \frac{1}{2}p_{1A}\right) - \left(p_{2U} + \frac{1}{2}p_{1U}\right)\right]^2}{\left(\frac{1}{n_{.A}} + \frac{1}{n_{.U}}\right)\left(\frac{1}{n_{..}^2}\right)\left[n_{..}\left(\frac{1}{4}n_{1.} + n_{2.}\right) - \left(\frac{1}{2}n_{1.} + n_{2.}\right)^2\right]}$$

where

$$p_{ij} = \frac{n_{ij}}{n_{.j}}$$

- X^2 has χ^2 distribution with 1 degree of freedom under null hypothesis.

	Cases	Controls	Total
MM	n_{2A}	n_{2U}	$n_{2.}$
Mm	n_{1A}	n_{1U}	$n_{1.}$
mm	n_{0A}	n_{0U}	$n_{0.}$
Total	$n_{.A}$	$n_{.U}$	$n_{..}$

- Odds ratio for allele M relative to allele m

$$\psi_{M|m} = \frac{\left(\frac{n_{1A}n_{0U}}{n_{0.}+n_{1.}}\right) + \left(\frac{n_{2A}n_{1U}}{n_{1.}+n_{2.}}\right) + \left(\frac{4n_{2A}n_{0U}}{n_{0.}+n_{2.}}\right)}{\left(\frac{n_{0A}n_{1U}}{n_{0.}+n_{1.}}\right) + \left(\frac{n_{1A}n_{2U}}{n_{1.}+n_{2.}}\right) + \left(\frac{4\left[n_{2A}n_{2U}n_{0A}n_{0U}\right]^{1/2}}{n_{0.}+n_{2.}}\right)}$$

- Affected individual $\psi_{M|m}^2$ times more likely to have marker genotype MM than mm, and $\psi_{M|m}$ times more likely to have genotype Mm than mm.

Allele-based Single-locus Tests

- Each individual now contributes ***two*** counts to the contingency table, one for each allele in their marker genotype.
- Assuming the sample to be typed at a marker SNP of interest, we can represent genotype data in a 2 x 2 contingency table.
- To test the null hypothesis of no disease-marker association
- Where $$X^2 = \sum_{i=0,1} \sum_{j=A,U} \frac{\left(n_{ij} - E\left[n_{ij}\right]\right)^2}{E\left[n_{ij}\right]}$$

$$E\left[n_{ij}\right] = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}$$

- X^2 has χ^2 distribution with 1 degree of freedom under null hypothesis.

	Cases	Controls	Total
M	n_{1A}	n_{1U}	$n_{1\cdot}$
m	n_{0A}	n_{0U}	$n_{0\cdot}$
Total	$n_{\cdot A}$	$n_{\cdot U}$	$n_{\cdot\cdot}$

- Odds ratio for allele M relative to m

$$\psi_{M|m} = n_{1A} n_{0U} \big/ n_{0A} n_{1U}$$

- Allele M is $\psi_{M|m}$ times more likely to be carried by an affected individual than allele m.
- Assumes multiplicative disease risks and Hardy-Weinberg equilibrium at SNP in cases and controls.

Unknown Phase in Multilocus Analysis

- Haplotype is a collection of alleles from neighbouring loci
- There are $2^{(L-1)}$ possible phases if L loci are heterozygous
- Let $\theta = (\theta_1, \theta_2, \ldots, \theta_J)$ be the vector of haplotype frequencies. Assuming Hardy-Weinberg equilibrium,

$$P(g) = P(h) = \prod_{j-1}^{J} \binom{2}{h_j} \theta^{h_j} = \binom{2}{h_j} \prod_{j-1}^{J-1} \binom{2}{h_j} \varphi^{h_j} \left(1 + \sum_{j=1}^{J-1} \varphi_j\right)^{-1}$$

and

$$\varphi_j = \theta_j / 1 - \sum_{j=1}^{J-1} \theta_j$$

Table 1
Genotype counts for biallelic markers

Marker 1	Marker 2		
	1/1	1/2	2/2
1/1	n_0	n_1	n_2
1/2	n_3	n_4	n_5
2/2	n_6	n_7	n_8

Table 2
Genotypic probabilities for two biallelic markers

Marker 1	Marker 2		
	1/1	1/2	2/2
1/1	h_{11}^2	$2h_{11}h_{12}$	h_{12}^2
1/2	$2h_{21}h_{11}$	$2(h_{21}h_{12}+h_{22}h_{11})$	$2h_{22}h_{12}$
2/2	h_{21}^2	$2h_{21}h_{22}$	h_{22}^2

For three SNPs, we have a very nice cube to demonstrate when the heterozygote(s) have to be considered

Zhao and Sham (2003) *CMPB*

The Combined Model

The complete data log-likelihood for the *i*th individual ignoring some constants is given by LMM and more general GLMM

$$l_i = \frac{y_i \eta - b(\eta)}{a(\phi)} + \sum_{j=1}^{J-1} h_{ij} \log \varphi_j - 2\log(1 + \sum_{j=1}^{J-1} \varphi_j)$$

where $\eta = X\beta + Z\gamma$. If t_i indicates both genetic and environmental effects, then an EM algorithm involves,

$$P(h_{ij} h_{ij'} \mid d_i^{(o)}) = \frac{f(y_i \mid t_i, z_i) p(g_i)}{\sum_{g \in S(G)} f(y_i \mid t_i, z_i) p(g_i)}$$

Lake et al. (2003) Hum Hered

Transmission Disequilibrium Test (TDT)

- For a large sample of trios (with varying genotypes) we can form a table. Under the null hypothesis, transmissions of allele 1 from a 1/2 parent are equally likely as transmissions of allele 2, i.e. we expect *b* = c. The test of association is McNemar's: $(b - c)^2/[b + c]$ with chi-squared distribution, (b + c) is an estimate of the variance of $(b-c)^2$; so that data from only 2 cells (b, c) are used.
- Transmissions from homozygous parents and (1/1 or 2/2) are ignored in the analysis.
- Alternative, more powerful statistic such as HHRR, but is not protected against population stratification
- It has been extended to multiallelic markers, quantitative traits, pedigree data and only siblings.

	U		
T		1	2
	1	a	b
	2	c	d

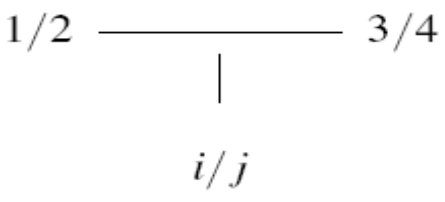
T = transmitted allele

U = untransmitted allele

The relative risk estimate (RR) b/c with variance estimate
Var(log(RR)) = 1/b + 1/c

Conditional on Parental Genotypes

- Consider parents with genotypes $1/2$ and $3/4$. An unselected offspring can have one of four genotypes with equal probability:



$$i/j = \begin{matrix} 1/3 \\ 1/4 \\ 2/3 \\ 2/4 \end{matrix} \quad \text{each with probability } 0.25$$

- Denoting the risk of disease in a subject with genotype i/j by $\pi\theta_{i/j}$, where θ denotes a *genotype relative risk* then, if we choose the family with an *affected* offspring, then the probability that the offspring has genotype i/j is

$$\frac{\theta_{i/j}}{\theta_{1/3} + \theta_{1/4} + \theta_{2/3} + \theta_{2/4}}$$

- Writing $\theta_{i/j} = \phi_i\phi_j$, the conditional probability becomes

$$\frac{\phi_i\phi_j}{\phi_1\phi_3 + \phi_1\phi_4 + \phi_2\phi_3 + \phi_2\phi_4} = \frac{\phi_i}{\phi_1 + \phi_2} \times \frac{\phi_j}{\phi_3 + \phi_4}$$

Bayesian vs Classical Statistics

Models $P(Data|\theta_1, M_1)$ $P(Data|\theta_2, M_2)$

Priors $P(\theta_1|M_1)$ $P(\theta_2|M_2)$

$$\text{Bayes Factor} = \frac{P(Data|M_1)}{P(Data|M_2)} = \frac{\int P(Data|\theta_1, M_1)P(\theta_1|M_1)d\theta}{\int P(Data|\theta_2, M_2)P(\theta_2|M_2)d\theta}$$

$P(Data|M_1)$ and $P(Data|M_2)$ are marginal likelihoods.

$$\text{Likelihood Ratio} = \frac{\max_{\theta_1} P(Data|\theta_1, M_1)}{\max_{\theta_2} P(Data|\theta_2, M_2)}$$

BF is interpreted as the factor by which the prior odds of association are changed in light of the data to produce the posterior odds of association: Posterior odds=BF x Prior odds. For instance M_1 denotes the model in which the SNP is associated with an additive effect on the log-odds scale, in contrast to the null model M_0 of no association, then $\theta_1=(\mu,\gamma)$, $\log(p_i/(1-p_i)= \mu+ \gamma Z_i$, and $\theta_0=(\mu,\gamma)$, $\log(p_i/(1-p_i)= \mu$. We can take $\mu \sim N(0,1)$, with N(.,.) for γ.

Marchini et al. (2007) Nat Genet

Bayes Factor

- Given the prior effect size as $N(0,\sigma^2)$
- $BF = (1+c^2)^{-0.5}\exp(z^2/2(1+c^2))$ with $c = \sigma/s$
  - P depends on z
  - BF depends on s, therefore sample size
  - We expect most null hypotheses to be true, i.e., the prior odds are against association, therefore a large BF is required and a small p-value, a compromise has been suggested to be when $N \approx 1/\sigma^2$, namely N is proportional to $1/(\text{effect size})^2$
- In GWAS the prior odds against association for any gene might be approximately 3,000:1, requiring a BF of at least 10^4, this corresponds to p–values $<10^{-6}$

Software

- POWER, a general package for power calculation
- QUANTO, power calculation for gene, environment and G-E interaction
- ANCESTRYMAP, admixture mapping (also ADMIXMAP but MALDsoft)
- 2LD/GENECOUNTING/HAP, programs for linkage disequilibrium analysis
- HAPLOVIEW, haplotype estimation, visualisation, tagging
- SNPHAP, haplotype estimation involving SNPs
- PHASE, haplotype inference using MCMC and coalescence
- tagSNPs, SNP tagging including generation of SAS program for haplotype trend regression
- FBAT, family-based association tests
- TRANSMIT, transmission-disequilibrium test (TDT) with haplotypes
- UNPHASED, haplotype analysis involving unrelated individuals or families
- HelixTree, GENOMIZER, PLINK, SNPGWA, tools for GWAS

Genomewide Association Studies

- Large Data
- Multiple Testing
- Graphical Methods
- HapMap
- Example Use

Large Data

- Database management systems might be required, as may be already in place in some epidemiological cohorts
- If a SNP is stored as 1 byte, for a 1M GeneChip and 10,000 individuals one needs 10GB storage. One may not be able to click to examine the raw data
- We can employ specialized programs for standard analysis
- General systems can be viable for analysis of a range of genetic and nongenetic factors

Multiple Testing

- The false discovery rate measures the proportion of false positives among all SNPs called significant:

$$FDR = \left[\frac{F}{S} \mid S > 0\right] p(S > 0) \qquad \text{or} \qquad pFDR = E\left[\frac{F}{S} \mid S > 0\right]$$

  where S is the total number of tests called significant but only F of which are true. The p-value is a measure of significance in terms of false positive rate (Type I error rate). The *q*-value is an FDR measure of significance associate with each SNP.

- Since $q - value(p_i) = \min_{t \geq p_i} pFDR(t) \ \forall t \in [0,1]$ and for a large number of SNPs, we can use FDR for the estimate of *q*-value. The parameter λ is chosen to tune for the proportion of truly null π_0 to obtain the FDR estimate or approximately *q*-value.

- The problem recently has largely been gotten around using a threshold of genomewide significance plus replication.

The Prior Probability of Linkage and FDR

- Morton (1955) set the prior prbability of linkage $\pi \approx 1/20$, for a given type I error rate α, and power W, we have

$$FDR = \frac{\alpha(1-\pi)}{\alpha(1-\pi)+\pi(1-\beta)} = \frac{10\alpha}{19\alpha+(1-\beta)}$$

- Smith (1953) assumed $\pi = 1/24, \alpha = 0.05, 1-\beta = 1$,

$$FDR = (23/20)(23/20+1) = 0.535$$

- The relationship between type I error α and power $1-\beta$. Since $\alpha = \int_C f_0(x)dx, 1-\beta = \int_C f_1(x)dx$, we have $\alpha/(1-\beta) = \int_C f_0(x)/\int_C f_1(x)dx = \int_C (f_0(x)/f_1(x))f_1(x)dx/\int_C f_1(x)dx = E_1(f_0(x)/f_1(x)|x \in C)$. So if C is region $f_0/f_1 \leq 1/A$, then $\alpha/(1-\beta) \leq 1/A$. For the SPRT, we set $\alpha/(1-\beta) = 1/A$ or $A = (1-\beta)/\alpha$ and $B = \beta/(1-\alpha)$

- These are linked with FPRP of Wachholder et al. (2004) and BFDP of Wakefield (2007)

Data Pre-processing

- It can take up more time than statistical analysis, big datasets which is neither possible to sort, nor transpose, nor analyse, data are partitioned to avoid specification of individual variable names, but
  - There might be difficulty with variable types, e.g., a mixture of numeric, character
  - If it is through customised code, do we take good care of the abnormalities?
- We might only need information from a subset
- SAS
  - Can handle large data, be good at merging and data partitioning, be used to produce input/output for other programs
  - Extensive modelling ability
- awk
  - Very powerful and good at jobs which may turn to be difficult for SAS, Stata or R, Pretty light and without need for sophisticated tools such as Perl
  - Readily available from both Unix/Linux and Windows systems and well-documented in the Internet

Zhao et al. (2007) CCIS

The Wide – This Is the Familiar and the Standard

ID	Age	Sex	BMI	SNP1
WTCCC139236	60	2	34.135406494	A/C
WTCCC139239	60	2	31.002960205	CC
WTCCC139240	50	2	30.036514282	AA
...				

The long – This Is the Convenient and the Supported

ID	rsn	Pos	age	sex	BMI	Add	Dom	Rec
1	snp1	1231	60	1	30	0	0	0
1	snp2	7891	60	1	30	1	1	0
1	snp3	12321	60	1	30	2	1	1
2	snp1	12331	30	2	35	2	1	1
2	snp2	15312	30	2	35	1	1	0
2	snp3	22312	30	2	35	0	0	0
…								

The Flipped – e.g. HapMap

Chr	rsn	Position	ID1	ID2	…
1	snp1	123	A/C	CC	
1	snp2	223	C/T	TT	
1	snp3	323	A/G	GG	
…					
2	snp1	100	C/G	GG	
2	snp2	200	GG	A/G	
…					

The Imputed: Genotypes Are Probabilities

chr	rsn	Position	Allele A	Allele B	Genotypes of ID1		
1	snp1	1245	A	C	0.00	0.12	0.88
1	snp2	1260	C	T	0.80	0.20	0.00
...							

Conversions

- They may be required in a variety of scenarios
- They can be done to a different degrees by
  - PROC TRANSPOSE/IML (**SAS**)
  - FLIP (**SPSS**)
  - reshape (**Stata**, **R**)
  - awk/gawk/nawk, ruby, Perl, TCL/TK, Java
  - **LINKAGE**, MERLIN, HaploView, GTOOL, PLINK, BC/SNPmax?
  - C/C++/C#, Fortran

More on Data Format

- The long and skinny format duplicates information.
- The wide format is familiar but requires specification of SNP names, or use of macro function in the case of SAS.
- The transposed format will be the one to go, with the expectation of a larger data/information base, e.g., the 1000 genome project. Perhaps the major benefit would be the ease with data extraction.
- Or probably it is not a good idea to store the imputed data after all, we can have SAS to grab the data online.

Allele-coding not a Big Deal!

Table 1. Allelic coding when the minor allele A is coded as B by alphabetical order

Correct Model	Genotype coding			Coded Model	Genotype coding			Change direction of effect
	A/A	A/B	B/B		A/A	A/B	B/B	
Additive	2	1	0	Additive	0	1	2	Yes
Dominant	1	1	0	Recessive	0	0	1	Yes
Recessive	1	0	0	Dominant	0	1	1	Yes

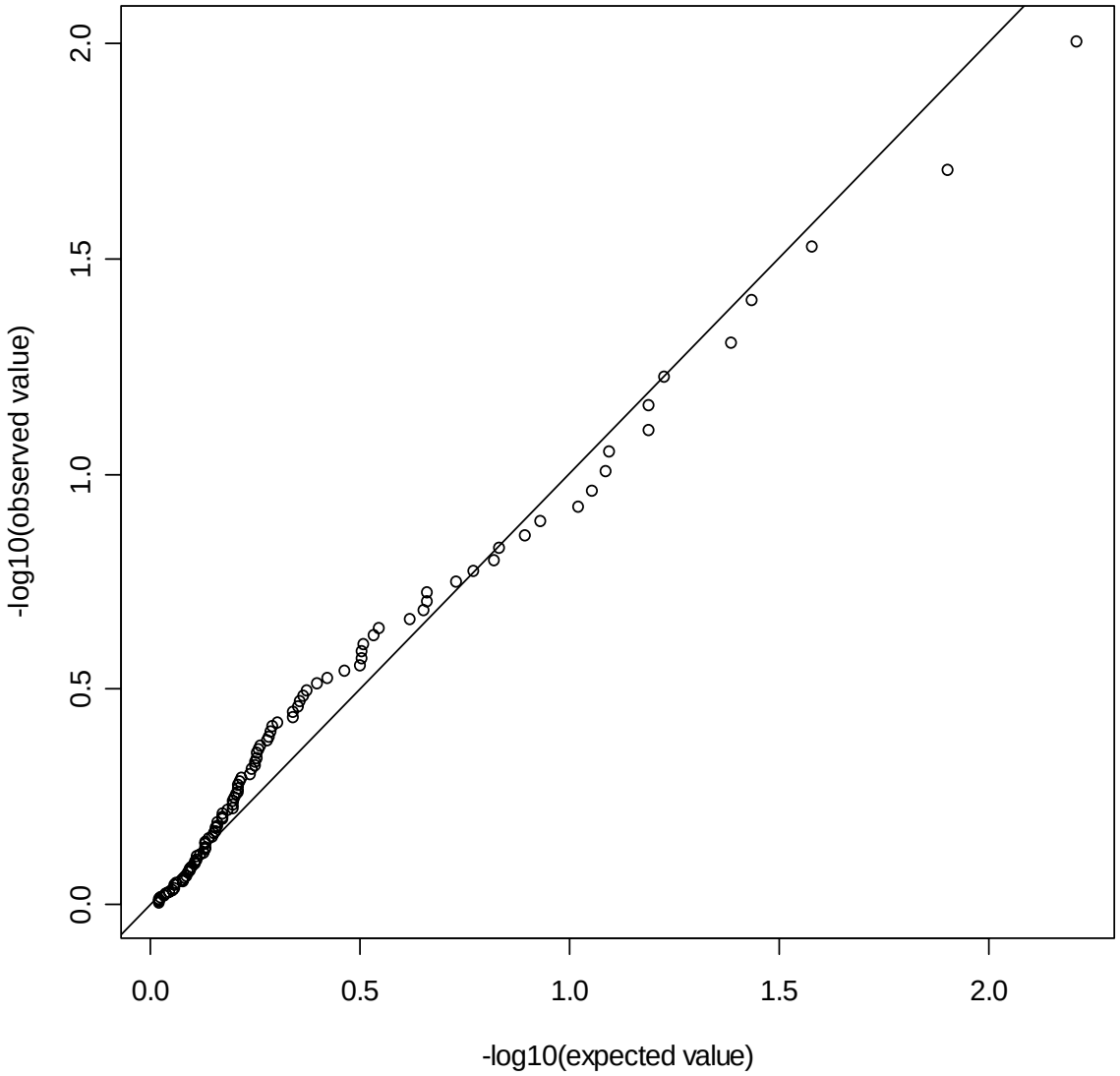
The bottleneck has been allele-coding, but the inclusion of map information would do away with it.

Graphical methods

- The plot of the observed versus the expected values
- Commonly used for p values

```
library(gap)
p <- runif(100)
r <- qqunif(p)
```

See Davison AC (2003) Statistical Models

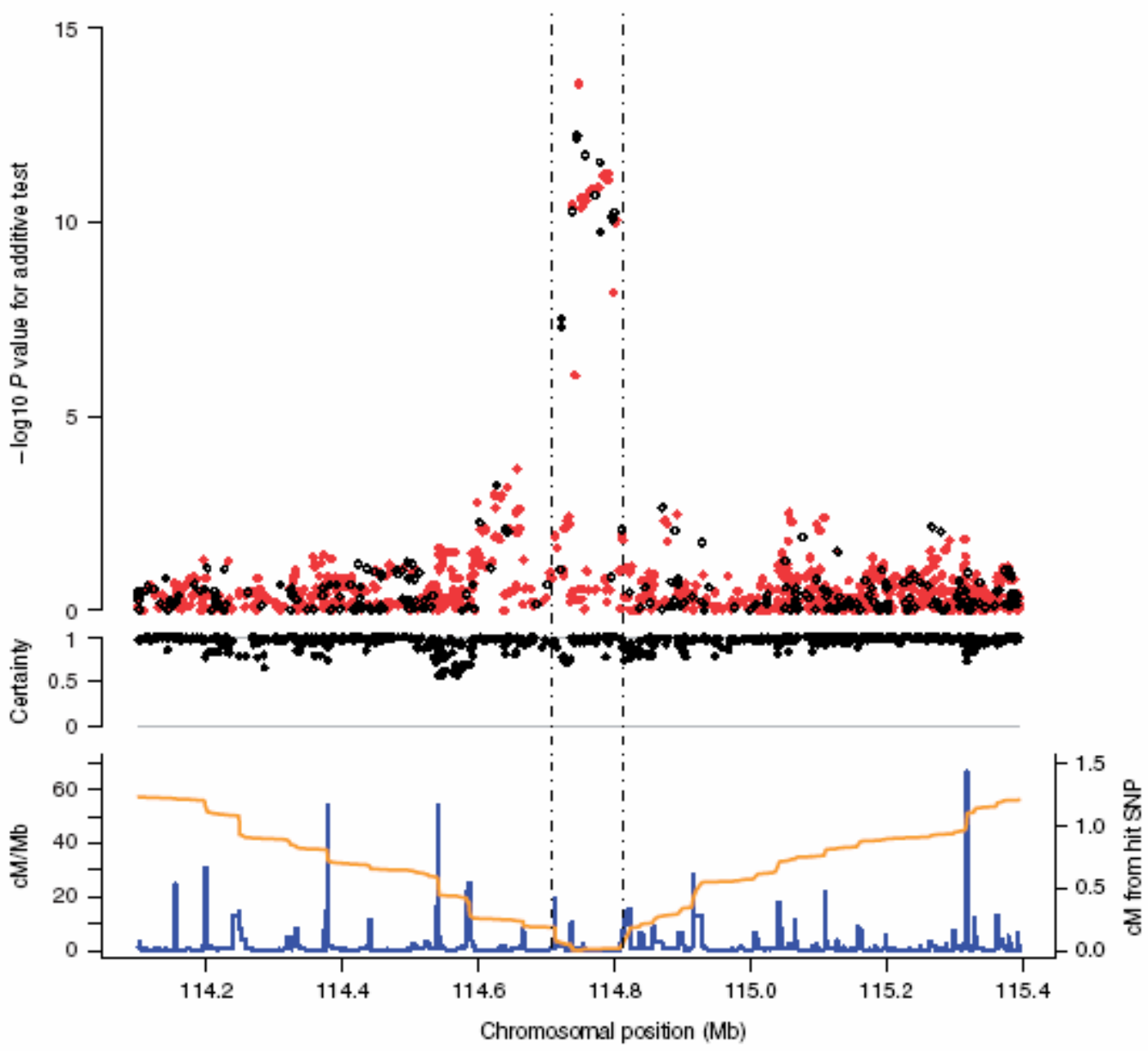


The HapMap Project

- The goal of the International HapMap Project is to develop a haplotype map of the human genome, the HapMap, which will describe the common patterns of human DNA sequence variation. The HapMap is expected to be a key resource for researchers to use to find genes affecting health, disease, and responses to drugs and environmental factors. The information produced by the Project will be made freely available.
- The project is a collaboration among scientists in Japan, the U.K., Canada, China, Nigeria, and the U.S. The Project officially started with a meeting on October 27-29, 2002 (http://genome.gov/10005336)
- It has been used for
  - Examine LD structure
  - Infer untyped genotypes (imputation)

Manolio et al. (2008) *J Clin Invest* 118:1590-1605

T2D hit region, chromosome 10



Meta-analysis

- Meta-analysis based on Fisher's p values
- Fixed and random effects model
- In simple terms, the former can be seen as inverse variance weighted estimate for significant testing, while the latter takes into account the heterogeneity between studies
- However, the heterogeneity is always worthwhile checking than fixed or random effects modelling

Normand (1999) *Stat Med*, van Houwelinggen et al. (2002) *Stat Med*

The Virtue of Meta-analysis

```
np <- 7
p <- 0.1^((np+1):2)
z <- qnorm(1-p/2)
n <- c(32000,8000)
n1 <- n[1]
s1 <- s2 <- vector("numeric")
for (i in 1:np)
{
   a <- z[i]
   for (j in 1:np)
   {
      b <- z[j]
      metaz1 <- (sqrt(n1)*a+sqrt(n[1])*b)/sqrt(n1+n[1])
      metap1 <- pnorm(-abs(metaz1))
      metaz2 <- (sqrt(n1)*a+sqrt(n[2])*b)/sqrt(n1+n[2])
      metap2 <- pnorm(-abs(metaz2))
      k <- (i-1)*np+j
      cat(k,"\t",p[i],"\t",p[j],"\t",metap1, "\n")
      s1[k] <- metap1
      s2[k] <- metap2
  }
}
```

```
log10 <- function (x) log(x,10)
q <- -log10(sort(p,decreasing=TRUE))
t1 <- matrix(-log10(sort(s1,decreasing=TRUE)),np,np)
t2 <- matrix(-log10(sort(s2,decreasing=TRUE)),np,np)

par(mfrow=c(1,2),bg="white",mar=c(4.2,3.8,0.2,0.2))

…
```

The p-values are one-sided if both betas have same directions

For instance, n=3,200 and p=1E-6, the combined p=2.65E-16; and even with p=1E-4 and 0.01, the combined p=2.41E-6

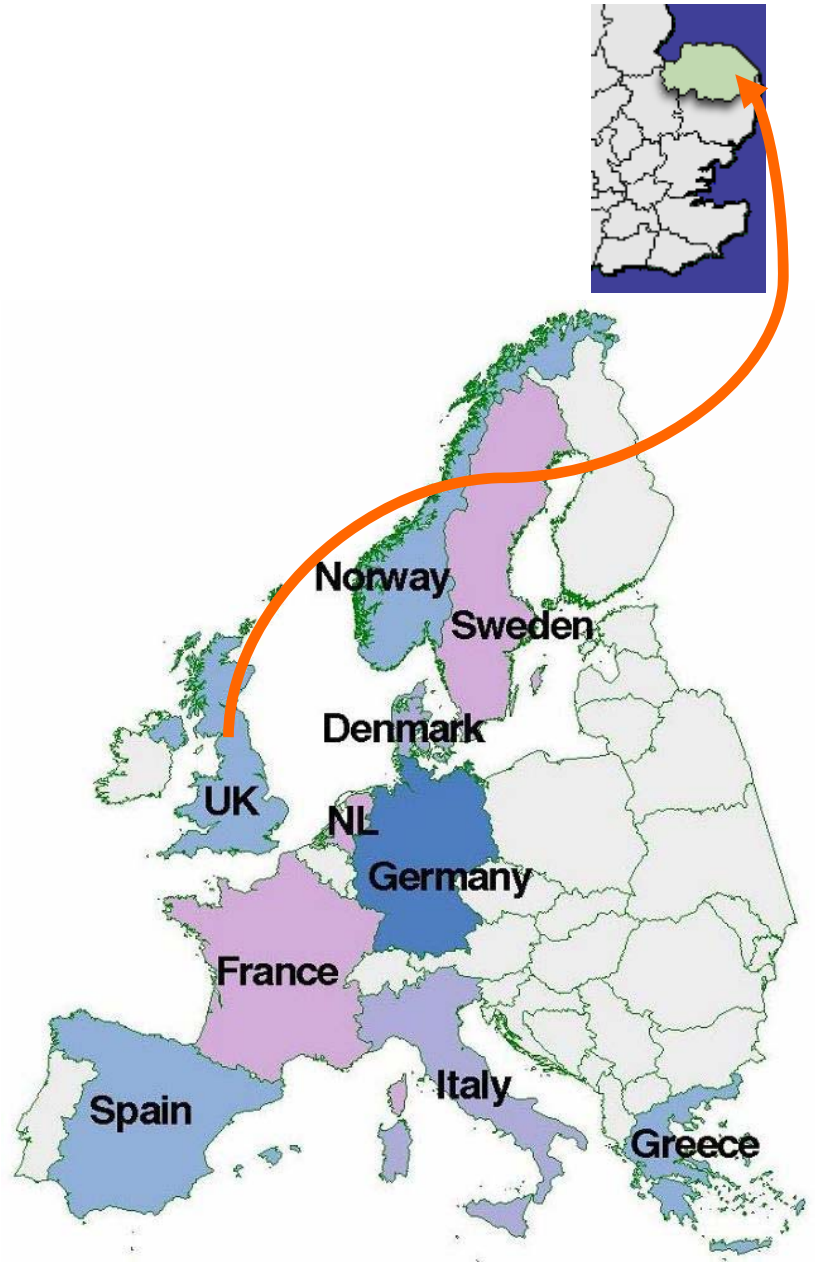
EPIC study

The European Prospective Investigation into Cancer and Nutrition (EPIC) is coordinated by Dr Elio Riboli, Head of the Division of Epidemiology, Public Health and Primary Care at the Imperial College London.

EPIC was designed to investigate the relationships between diet, nutritional status, lifestyle and environmental factors and the incidence of cancer and other chronic diseases. EPIC is the largest study of diet and health ever undertaken, having recruited over half a million (520,000) people in ten European countries: Denmark, France, Germany, Greece, Italy, The Netherlands, Norway, Spain, Sweden and the United Kingdom.

EPIC-Norfolk study

EPIC-Norfolk participants are men and women (based on over 30,000 people) who were aged between 45 and 74 when they joined the study, who lived in Norwich and the surrounding towns and rural areas. They have been contributing information about their diet, lifestyle and health through questionnaires, and through health checks carried out by EPIC nurses.



Case-cohort design for EPIC-Norfolk study

- It originally followed case-control design (e.g., WTCCC with seven cases and common controls) with 3425 cases and 3400 controls.
  - It is potentially more powerful.
  - Controls are selected.
- It has then been changed into case-cohort design, in which cases are defined to be individuals whose BMI above 30 and controls are a random sample (subcohort) of the EPIC-Norfolk cohort which includes obese individuals.
  - The subcohort is representative of the whole population and allows for a range of traits to be examined.
  - The analysis is potentially more involved but established.

Power/sample size

- It started with assessment of how the power is compromised relative to the original case-control design.
- This was followed by power/sample size calculation using methods established by Cai and Zeng (2004) as implemented in an R function, noting a number of assumptions.
  - The censoring distributions are the same in the two groups
  - The number of failures is very small in the full cohort but much larger than one
  - No ties of failures are observed
- More practically, it was also envisaged that a proper representative sample of a total of 25,000 individuals would be 10%; the subcohort is then approximately 2,500.
- The total sample was split between two stages.

GeneChips

- Affymetrix 500K
  - Data were available for 3850 individuals
- Illumina 317K
  - It came at a later time。
  - Data quality appears to be poor?
- The focus has therefore been Affy500K, but with a possible comeback.

Analysis

- An incremental approach was adopted since the storage and computing power were somewhat uncertain.
- This was predated with controls from the breast cancer study, involving about 400 individuals with Perlegen 250K GeneChips.
- QC including call rates and HWE was feasible with SAS/Genetics (~30GB) which provides a good estimate of the storage for all individuals (~380GB).
- The Linux platform seemed to be favourable.

EPIC400 analysis

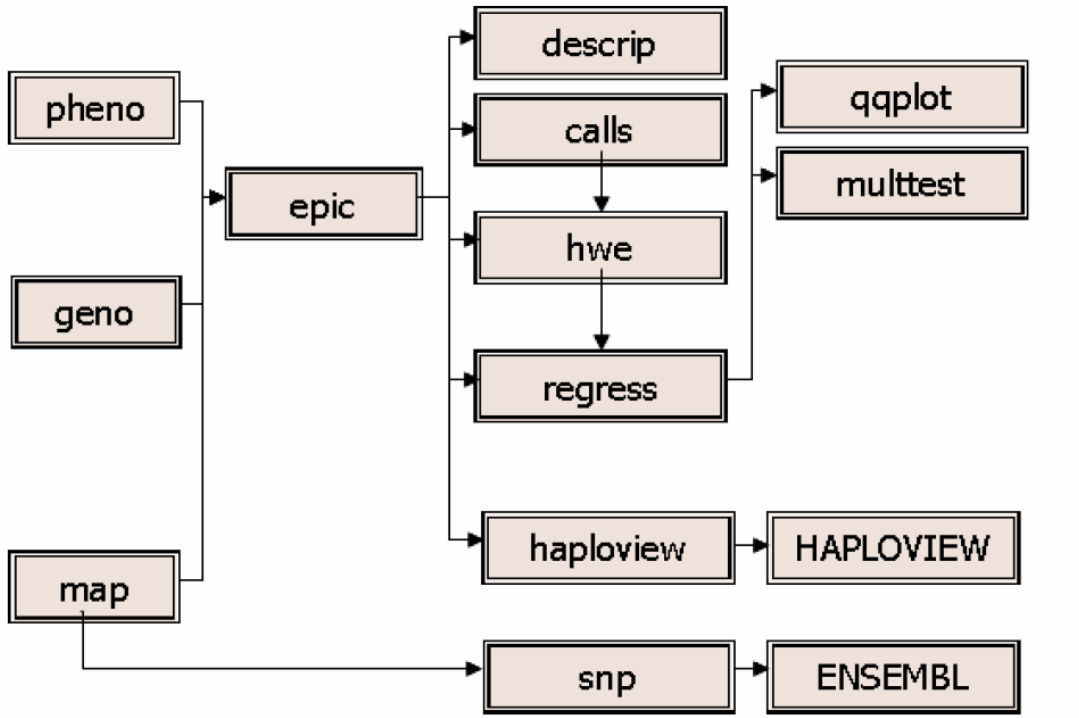


Fig. 1. A flowchart of the EPIC 400 analysis, with modules in brackets. Genotypes (geno) and phenotypes (pheno) are merged (epic) for descriptive statistics (descrip) call rates (calls), HWE (hwe), regression (regress) with adjustment for multiple testing (multtest) and comparison with theoretical distribution (qqplot). The raw data together with map information (map) can also be reformatted (haploview) into HAPLOVIEW input files so that specific region in the genome can be visualized, with annotation information from ENSEMBL according to SNPs (snp).

The analysis for GWAS

- QC including visualisation of clustering, outliers, was largely done by colleagues at Sanger (as for WTCCC): call rates, clustering, duplicates, Hardy-Weinberg equilibrium, minor allele frequencies (MAF), agreement with HapMap, ethnic outliers
- The overall strategy was data partition, i.e., by chromosome and further by region (30) in each chromosome, largely on a long, skinny data format
- A major advantage is that the analysis can be resumed whenever the system experiences problems
- We stuck to SAS to allow for reliability and flexibility with or without SAS/Genetics, for BMI/obesity as continuous and binary outcomes are readily tackled with REG/LOGISTIC procedures – most outputs are available from the output delivery system (ODS)
- The picture was eventually changed with a revised coding algorithm and the use of imputed data

Additional analysis

- Population stratification via EIGENSTRAT
  - SAS is very handy since a single put statement is sufficient to generate the output.
- Collaborative (e.g. height) and consortium work (GIANT)
  - On the UK side, this is mainly involved with IMPUTE/SNPTEST, with inputs on strand, standard error, quantitative traits, outputs.
  - This facilitates meta-analysis considerably.

Population Stratification (EIGENSTRUCT)

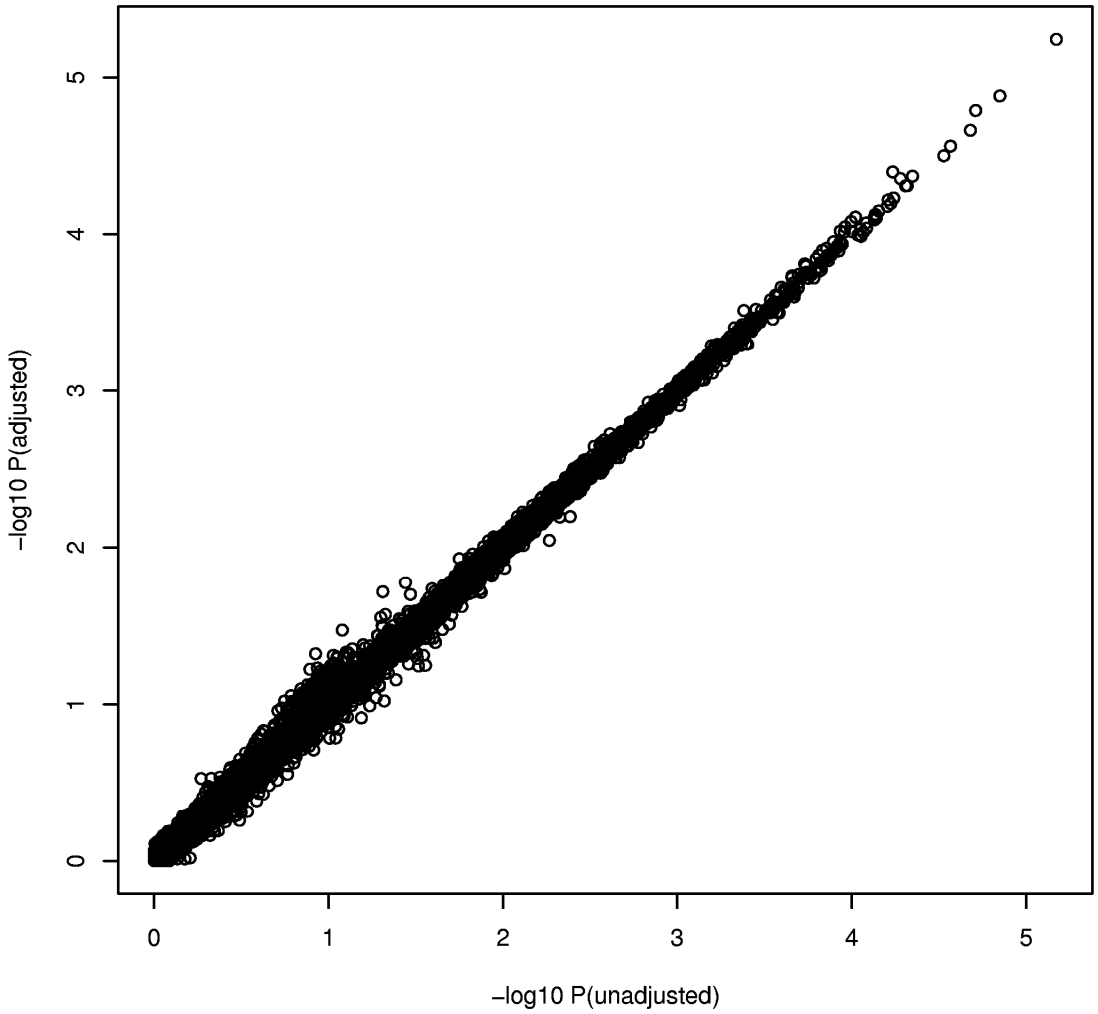
```
x2 <- read.table("epic5k.chisqQTL",skip=7,header=TRUE)
attach(x2)
p1 <- 1 - pchisq(Chisq,1)
p2 <- 1 - pchisq(EIGENSTRAT,1)
logp1 <- -log(p1,10)
logp2 <- -log(p2,10)
bitmap("logp.bmp",res=72*5)
plot(logp1,logp2,type="p",
xlab="-log10 P(unadjusted)",
ylab="-log10 P(adjusted)")
hits <- p1<5e-7
points(logp1[hits],logp2[hits],col="g
dev.off()
```

The result for height, lambda=1.04

And the top ten eigenvalues are

2.98, 2.86, 2.79, 2.67,

2.66, 2.43, 2.31, 2.25, 2.18, 2.13



Single-point analysis

- Linear regression
  - Appropriate for BMI, LDL
  - Good estimate for variance%

- Logistic regression
  - Appropriate for binary trait such as obesity
  - Availability of OR, PAR estimates

- Implemented in SAS and R

- Truncated regression experiment with PROC QLIM

Haplotype Analysis

- We focus on regions of interest to reduce the computational burden, with the S-PLUS/R package haplo.stats

```
library(haplo.stats)
locus.label <- c("rs17782313", "rs12955983", "rs17700633",
    "rs718475", "rs9956279")
haplo.score(zl10bmi,snps,x.adj=age,locus.label=locus.label,simulate
    =TRUE)
```

- This code will perform haplotype analysis of phenotype zl10bmi, with object snps containing five SNPs, adjusted for age and obtaining both actual and simulated p values

Haplotype Analysis

Global Score Statistics, global-stat = 34.20497, df = 17, p-val = 0.0079
Global Simulation p-value Results, Global sim. p-val = 0.02827, Max-Stat sim. p-val = 0.051
Number of Simulations, Global: 1026 , Max-Stat: 1000
Haplotype-specific Scores (rs17782313 rs12955983 rs17700633 rs718475 rs9956279)

Haplotype	Freq	Score	p-val	sim p-val
1, 2 1 2 2 1	0.57608	-2.80522	0.00503	0.009
2, 2 1 2 1 1	0.00799	-0.98884	0.32274	0.359
3, 1 2 2 1 2	0.00032	-0.97033	0.33188	0.332
4, 1 2 1 2 2	0.00076	-0.76784	0.44258	0.456
5, 2 2 2 2 1	0.01627	-0.73352	0.46324	0.454
6, 2 2 1 1 2	0.02727	-0.65671	0.51137	0.536
7, 2 1 1 1 2	0.12879	-0.54604	0.58504	0.57
8, 1 1 1 1 2	3e-04	0.21458	0.8301	0.829
9, 2 1 2 1 2	0.00853	0.31731	0.75101	0.751
10, 1 1 2 2 1	0.00812	0.61806	0.53654	0.512
11, 1 2 1 1 1	0.00028	0.74632	0.45548	0.481
12, 2 1 2 2 2	0.00133	1.08587	0.27753	0.277
13, 2 1 1 1 1	0.00059	1.34426	0.17887	0.18
14, 1 2 1 2 1	0.00078	1.36987	0.17073	0.17
15, 1 2 2 2 1	0.07891	1.88485	0.05945	0.045
16, 2 1 1 2 1	0.00071	2.95472	0.00313	0.005
17, 1 2 1 1 2	0.14233	3.20313	0.00136	0.002

Cross-check with PROC HAPLOTYPE

For the two-SNP case the code is as follows,

```
options nocenter;
libname x '.';
data test5;
        input snp1$ snp2$;
        a11=substr(snp1,1,1);
        a12=substr(snp1,2,1);
        a21=substr(snp2,1,1);
        a22=substr(snp2,2,1);
        if a11^="N" & a12^="N" then
           m1=compress(a11||"/"||a12);
        if a21^="N" & a22^="N" then
           m2=compress(a21||"/"||a22);
cards;
CT  AG
TT  GG
CT  AG
CT  AG
TT  AG
run;
proc print data=test5;;
run;
proc haplotype data=test5 genocol
        delimiter="/";
        var m1 m2;
run;
data chr18;
     set x.epicchr18_hap;
     a11=substr(rs17782313_add,1,1);
     a12=substr(rs17782313_add,2,1);
     a21=substr(rs17700633_add,1,1);
     a22=substr(rs17700633_add,2,1);
     if a11^="N" & a12^="N" then
        m1=compress(a11||"/"||a12);
     if a21^="N" & a21^="N" then
        m2=compress(a21||"/"||a22);
run;
proc haplotype data=chr18 genocol
        delimiter="/";
        var m1 m2;
run;
```

Imputation

- The HapMap data provide haplotype information for specific populations
- The genotyped data (e.g., from Affymetrix 500k GeneChips) can be seen as incomplete data
- We can impute the “missing” genotypes across the genome for *in silico* genotypes
- This can be achieved via a HMM as implemented in IMPUTE and MACH
- The output can be represented as probabilities to be used by purpose-written computer program such as SNPTEST
- Customised programs have been written for IMPUTE/SNPTEST (SAS) and MACH (GenABEL)
- The actually genotyped data can be analysed as if they were imputed, to be in line with the analysis of imputed data
- The process can be populated through different studies to allow for uniform and meta-analysis

SAS in Conjunction with Standalone Programs

- There are more problems than established systems but faster due to less overhead.
- By working closely with the developer, it led to more flexible and faster turnover
- All collaborators are informed over the changes and standard operating practice
- Specifically, we applied SNPTEST for both actually genotyped (inputs to IMPUTE) and imputed data in ~70 analyses for the EPIC-Norfolk data

Meta-analysis

- SAS macros were implemented
- Also available from Stata *(metan)* which generates the forest plot
- To facilitate easy use, a C/C++ program METAL was used, and validated with R
- It allows for SNPs unavailable from particular platforms (e.g., Affymetrix and Illumina) and QC criteria to be used
- The heterogeneity test is an invaluable source to study interaction

Genetic Investigation of Anthropometric Traits

- An international initiative created to study a range of anthropometric traits including BMI, height, waist
- Through comprehensive meta-analyses of summary statistics based on analysis of actually genotyped and imputed data
- Successful replication of top hits with GIANT and deCODE

Findings Regarding Obesity and LDL

- T. M. Frayling, N. J. Timpson, M. N. Weedon, E. Zeggini, R. M. Freathy, C. M. Lindgren, J. R. Perry, K. S. Elliott, H. Lango, N. W. Rayner, B. Shields, L. W. Harries, J. C. Barrett, S. Ellard, C. J. Groves, B. Knight, A. M. Patch, A. R. Ness, S. Ebrahim, D. A. Lawlor, S. M. Ring, Y. Ben-Shlomo, M. R. Jarvelin, U. Sovio, A. J. Bennett, D. Melzer, L. Ferrucci, R. J. Loos, I. Barroso, N. J. Wareham, F. Karpe, K. R. Owen, L. R. Cardon, M. Walker, G. A. Hitman, C. N. Palmer, A. S. Doney, A. D. Morris, G. D. Smith, A. T. Hattersley, and M. I. McCarthy. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 316 (5826):889-894, 2007.
- T. Gerken, C. A. Girard, Y. C. Tung, C. J. Webby, V. Saudek, K. S. Hewitson, G. S. Yeo, M. A. McDonough, S. Cunliffe, L. A. McNeill, J. Galvanovskis, P. Rorsman, P. Robins, X. Prieur, A. P. Coll, M. Ma, Z. Jovanovic, I. S. Farooqi, B. Sedgwick, I. Barroso, T. Lindahl, C. P. Ponting, F. M. Ashcroft, S. O'Rahilly, and C. J. Schofield. The obesity-associated FTO gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase. *Science* 318 (5855):1469-1472, 2007.
- M. S. Sandhu, D. M. Waterworth, S. L. Debenham, E. Wheeler, K. Papadakis, J. H. Zhao, K. Song, X. Yuan, T. Johnson, S. Ashford, M. Inouye, R. Luben, M. Sims, D. Hadley, W. McArdle, P. Barter, Y. A. Kesaniemi, R. W. Mahley, R. McPherson, S. M. Grundy, S. A. Bingham, K. T. Khaw, R. J. Loos, G. Waeber, I. Barroso, D. P. Strachan, P. Deloukas, P. Vollenweider, N. J. Wareham, and V. Mooser. LDL-cholesterol concentrations: a genome-wide association study. *Lancet* 371 (9611):483-491, 2008.

Empirical Power of the EPIC-Norfolk Case-Cohort Design

Sandhu et al. Lancet 2008

Supplementary Table S1. Associations between Affymetrix SNPs with a combined p-value of < 1.0×10^{-7} and circulating levels of LDL-c in independent study populations

SNP	EPIC-Norfolk sub-cohort		EPIC-Norfolk obese set		1958 British Birth Cohort		CoLaus		GEMs study	
	n = 2,269		n = 1,009		n = 1,375		n = 5,367		n = 1,665	
	β-coeff (se)	P-value	β-coeff (se)	P-value	β-coeff (se)	P-value	β-coeff (se)	P-value	β-coeff (se)	P-value
rs4420638	0.24 (0.04)	1.9×10^{-9}	0.14 (0.06)	0.02	0.25 (0.04)	2.8×10^{-9}	0.05 (0.01)	6.2×10^{-12}	0.04 (0.01)	5.6×10^{-3}
rs599839	-0.15 (0.04)	5.8×10^{-5}	-0.23 (0.06)	7.6×10^{-5}	-0.14 (0.04)	4.3×10^{-4}	-0.04 (0.01)	1.6×10^{-07}	-0.06 (0.01)	2.0×10^{-5}
rs4970834	-0.13 (0.04)	1.1×10^{-3}	-0.18 (0.06)	5.5×10^{-3}	-0.11 (0.04)	0.01	-0.04 (0.01)	1.9×10^{-06}	-0.04 (0.01)	2.8×10^{-3}
rs562338	-0.17 (0.04)	6.0×10^{-6}	-0.11 (0.06)	0.07	-0.18 (0.05)	1.1×10^{-4}	-0.03 (0.01)	2.7×10^{-06}	-0.02 (0.01)	0.18
rs7575840	0.15 (0.03)	6.3×10^{-6}	0.15 (0.05)	2.4×10^{-3}	0.04 (0.04)	0.26	0.03 (0.01)	1.9×10^{-06}	0.02 (0.01)	0.13
rs478442	-0.16 (0.04)	2.1×10^{-5}	-0.07 (0.06)	0.25	-0.16 (0.04)	3.6×10^{-4}	-0.03 (0.01)	2.7×10^{-5}	-0.02 (0.01)	0.06
rs4591370	-0.17 (0.04)	7.7×10^{-6}	-0.06 (0.06)	0.28	-0.16 (0.04)	4.2×10^{-4}	-0.03 (0.01)	3.2×10^{-5}	-0.02 (0.01)	0.06
rs4560142	-0.16 (0.04)	1.6×10^{-5}	-0.06 (0.06)	0.27	-0.16 (0.04)	4.2×10^{-4}	-0.03 (0.01)	3.5×10^{-5}	-0.03 (0.01)	0.05
rs576203	-0.16 (0.04)	1.2×10^{-5}	-0.07 (0.06)	0.25	-0.16 (0.04)	3.5×10^{-4}	-0.03 (0.01)	3.5×10^{-5}	-0.02 (0.01)	0.06
rs506585	-0.16 (0.04)	1.7×10^{-5}	-0.06 (0.06)	0.31	-0.16 (0.04)	3.5×10^{-4}	-0.03 (0.01)	4.2×10^{-5}	-0.03 (0.01)	0.05
rs488507	-0.14 (0.04)	1.3×10^{-4}	-0.07 (0.06)	0.25	-0.16 (0.04)	3.3×10^{-4}	-0.03 (0.01)	3.4×10^{-5}	-0.02 (0.01)	0.07
rs538928	-0.16 (0.04)	5.0×10^{-5}	-0.01 (0.06)	0.92	-0.16 (0.04)	3.5×10^{-4}	-0.03 (0.01)	3.6×10^{-5}	-0.02 (0.01)	0.05
rs10402271	0.04 (0.03)	0.17	0.11 (0.05)	0.02	0.12 (0.04)	7.5×10^{-4}	0.02 (0.01)	5.2×10^{-4}	0.04 (0.01)	8.3×10^{-4}

MC4R and Obesity

- Ruth Loos, Cecilia M. Lindgren, Shengxu Li, Ellie Wheeler, Jing Hua Zhao, Mike Inouye, Rachel M. Freathy, Antony Attwood, Jacques Beckmann, Sonja Berndt, Sven Bergmann, Amanda Bennett, Sheila Bingham, Murielle Bochud, Morris Brown, Stephane Cauchi, Cyrus Cooper, George Davey-Smith, Christian Dina, Subhajyoti De, Emmanouil Dermitzakis, Alex Doney, Kate Elliott, Paul Elliott, David Evans, Sadaf Farooqi, Philippe Froguel, Jilur Ghori, Christopher Groves, Rhian Gwilliam, David Hadley, Alistair Hall, Andrew Hattersley, Johannes Hebebrand, Iris Heid, Blanca Herrera, Sarah Hunt, Marjo-Riitta Jarvelin, Toby Johnson, Jennifer Jolley, Fredrik Karpe, Andrew Keniry, Kay-Tee Khaw, Robert Luben, Massimo Mangino, Jonathan Marchini, Wendy McArdle, Ralph McGinnis, David Meyre, Patricia Munroe, Andrew Morris, Andrew Ness, Matthew Neville, Alexandra Nica, Ken Ong, Stephen O'Rahilly, Katharine Owen, Colin Palmer, Konstantinos Papadakis, Simon Potter, Anneli Pouta, Lu Qi, Joshua Randall, William Rayner, Susan Ring, Manjinder Sandhu, Matthew Sims, Kijoung Song, Nicole Soranzo, Elizabeth Speliotes, Holly Syddall, Sarah Teichmann, Nicholas Timpson, Jonathan Tobias, Manuela Uda, Chris Wallace, Dawn Waterworth, Michael Weedon, Cristen Willer, Vicki Wraight, Xin Yuan, Eleftheria Zeggini, Joel Hirschhorn, David Strachan, Willem Ouwehand, Mark Caulfield, Nilesh Samani, Timothy M Frayling, Peter Vollenweider, Gerard Waeber, Vincent Mooser, Panos Deloukas, Mark McCarthy, Nicholas Wareham, Ines Barroso, Kevin Jacobs, Stephen Chanock, Richard Hayes, Claudia Lamina, Christian Gieger, Thomas Illig, Thomas Meitinger, H.-Erich Wichmann, Peter Kraft, Sue Hankinson, David Hunter, Frank Hu, Helen Lyon, Benjamin Voight, Martin Ridderstrale, Leif Groop, Paul Scheet, Serena Sanna, Goncalo Abecasis, Giuseppe Albai, Ramaiah Nagaraja, David Schlessinger, Anne Jackson, Jaakko Tuomilehto, Francis Collins, Michael Boehnke, and Karen Mohlke. Association studies involving over 90,000 samples demonstrate that common variants near to MC4R influence fat mass, weight and risk of obesity. Nature Genetics 2008.40:768-75

Ongoing and further work

- Gene-gene interaction (e.g., height)
- Covariate-SNP (e.g., gender, BMI) interaction
- Gene characterisation
- Family data, e.g., GAW16 and comments by Bodmer and Bonilla:

  "Family studies do not have a significant role in the discovery or analysis of either common or rare disease associated variants, both of which have relatively low penetrances at the individual level."
- Some work on R including gap, kinship, pan, useR!2008 (tutorials and graphical methods).

Specific analysis

- Which trait MC4R has effect on?
- Interpretation of mediation
  - Path analysis – shows mainly on BMI and not others
  - Error propagation as appropriate for meta-analysis

As is the case with FTO and T2D, the indirect effect (IE) from MC4R SNP to TG via BMI is $b_1 b_2$, with
$SE(IE) \approx b_1 SE(b_2)$

Reflection on the analysis

- A characterisation of the EPIC cohort? Maybe!
- an abstract to IGES 2007
  - BMI, or zBMI?
  - QLIM procedure, bivariate with SBP/DBP and HT
- What is the benefit of retrospective method?
- How about staged design?

The Advantages

- The benefit from large amount of genetic data outweighs the overhead.
- They have been shown to be highly successful in localising disease or trait association.
  - Diabetes, heart diseases, breast cancer
  - Obesity
  - Height
- It offers the possibility to examine or account for population differences and therefore allows for more international collaborations.

Genetic Association Study: Summary

- Association studies are the most powerful design
- Greatly facilitated by technological development
  - Computing
  - Genotyping and sequencing: RFLP, SSR, SNP
- Fine mapping
  - Molecular experiments
  - Statistical models
- More general framework
  - Multiple imputation
  - Meta-analysis

Retrospective Methods for Genetic Association

- Background
- Statistical Models
- A Summary
- More Details from Kwee et al

Background

- A paradigm of Cause ← Outcome
- Prentice RL, Pyke R. Biometrika 1979, **66**:403-11.
- The full retrospective likelihood is proportional to the retrospective likelihood of gene conditional on environmental factors and disease multiplied by the prospective likelihood of disease given genetic factors (Kwee et al. *GE* 2007, **31**:73-90). i.e., $L = P[G,E|D] = P[G|E,D]P[E|D]$
- TDT, Waldman et al. *AHG* 1999, **63**:329-40, Zou GY. *AHG* 2006, **70**:262-72
- Case-control design, e.g. Epstein MP, Satten GA. *AJHG* 2003, **73**:316-29; Weinberg CR, Umbach DM. *AJE* 2000, **152**:197-203, Power and bias were examined by Satten GA, Epstein MP, *GE* 2004, **27**:192-201
- Case-only design, e.g. Khoury & Flanners. *AJE* 1996, **144**:207-13
- Logistic model of discrete and continuous traits for population-based sample
  - Single-locus, Tan et al. *AHG* 2003, **67**:598-607;
  - Haplotype Analysis, Tan et al. *Genet Res* 2005, **86**:223-31
  - GEI, Tan et al. *BMC Genet* 2007, **8**:70
- To relax the independence assumption. Shin J-H, McNeney B, Graham J. SAGMB 2007, **6:**13; Chen et al. *Biostatistics* 2008, **9**:81-99

Statistical Models

Let

G = genetic variant, 0=non-carriers, 1=carriers

E = environment exposure, 0=non-exposed, 1=exposed

x = trait value

Then, we have

$Logit\ P[G=1,E=1|x]=a_G I+b_E J+(b_G I+b_E J+b_{GxE} IJ)x$

Where a=intercept (nuisance parameter), b=slope

We can write out $P[G=1,E=1|x]$, $P[G=0,E=0|x]$, $P[G=1,E=0|x]$ and therefore $RR_G(x)=exp(a_G+b_G x)$, $RRR_G(x_2:x_1)=exp[b_G(x_2-x_1)]$

The log-likelihood function is

$$l=\sum_{k=1}^{N}\sum_{i,j=0}^{1} I(G_k=I,E_k=J)p(a_G,a_E,b_G,b_E,b_{G\times E})$$

Summary

- A distinct feature of the model is trait is treated as an independent variable to allow for both categorical and continuous types and no assumption about normality is required (Cordell H. Hum Mol Genet 2002, **11**:2463-8; Hahn et al. Bioinformatics 2003, **19**:376-82).
- Supplementary materials showed that for the case of binary outcome, when the disease is rare, the relative risk estimate is comparable to that from a prospective model.
- It requires interaction variants be independent, otherwise a haplotype analysis model is required (Epstein MP, Satten GA. Am J Hum Genet 2003, **73**:1316-29; Tan et al. Genet Res 2005, **86**:223-31).
- It is equally applicable to study of any main and joint effects models

More Details from Kwee et al.

- A prospective likelihood approach by Lake et al. Hum Hered 2003, **55**:56-65 may suffer with respect to retrospective approach (Epstein & Satten 2004). Profile likelihood approaches by Lin et al. Genet Epidemiol 2005, 29:299-312; Spinka et al. Genet Epidemiol 2005, **29**:108-127 require estimating absolute risk of disease from case-control data
- Consider (a) rare disease and (b) haplotype-environment independence
- This leads to likelihood-based inference without specifying the distribution of environment covariates in the sample. Following the full likelihood specification, it is shown that $P[E|D]$ contained no information on haplotype and haplotype-environment interaction parameters. Furthermore, case-only study relies on information from $P[G|E,D=1]$
- Simulation studies showed that (with a realistic sample size) under a recessive model both the full retrospective and the prospective approaches yielded flawed results and occasional convergence problem. However, the multiplicative and dominant models give reasonable estimate.
- (To be) implemented in CHAPLIN for case-control haplotype analysis

Analysis of Pathways

- Statistical Methods
- Example Analyses
- A Case Study
  - Methods
  - Results
  - Summary

Statistical Methods

- Graphic models
  - WinBUGS
  - gR
- Bayesian networks
- Structural equation modelling
  - SPSS/AMOS, LISREL, Mx, EQS, MPlus, SYSTAT/EzPATH, SAS/CALIS, SAS/SYSLIN
  - R/sem, R/systemfit,

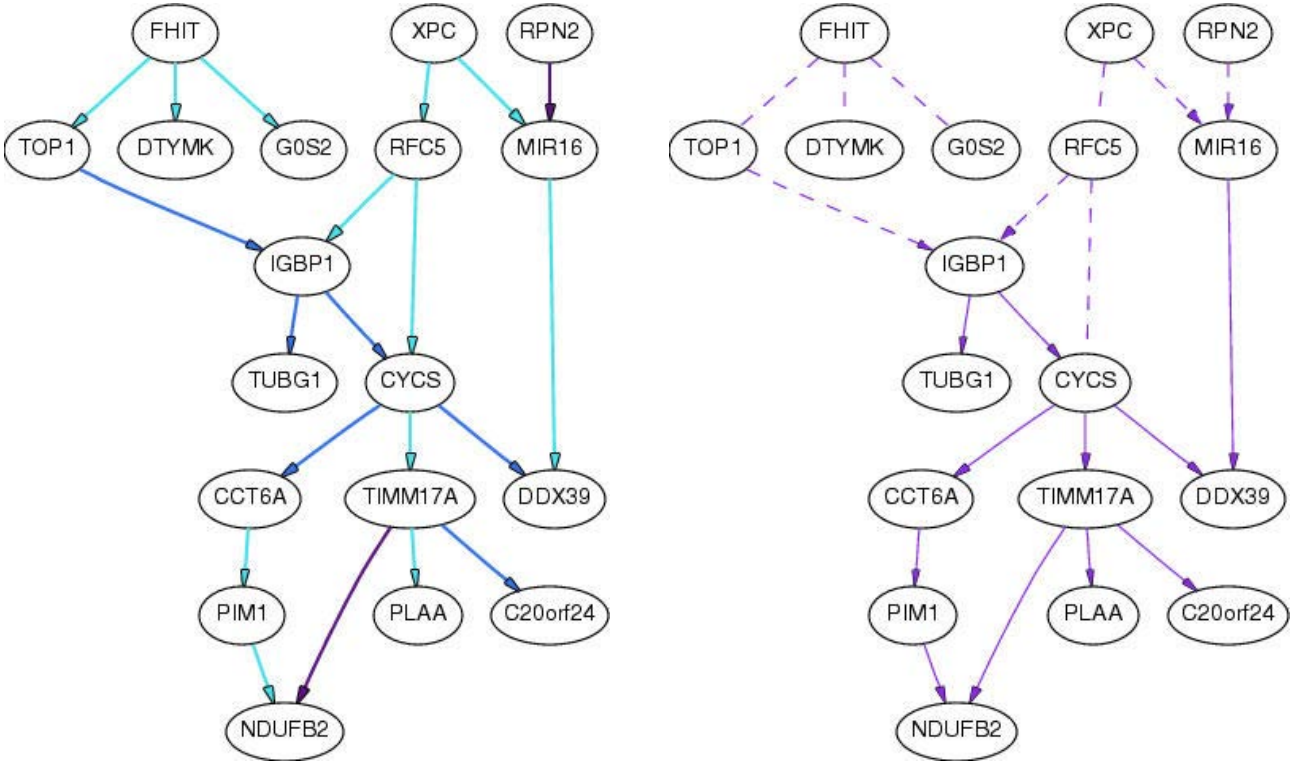
Bayesian Networks with GAW15 Problem 1

- Data from Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG: **Genetic analysis of genome-wide variation in human gene expression.** Nature 2004, **430:** 743-74
- Baseline expression levels of 8793 genes in immortalised B cells from 194 individuals in 14 Centre d'Etude du Polymorphisme Humane (CEPH) Utah pedigrees. Previous analysis of the data showed linkage and association and evidence of substantial individual variations.
- In particular, correlation was examined on expression levels of 31 genes and 25 target genes corresponding to two master regulatory regions
- In this analysis, we apply Bayesian network analysis to gain further insight into these findings. We identify strong dependences and therefore provide additional insight into the underlying relationships between the genes involved. More generally, the approach is expected to be applicable for integrated analysis of genes on biological pathways
- If the expression level of a given gene is regulated by certain proteins then it should be a function of the active levels of these proteins. Due to biological variability and measurement errors, the function would be stochastic rather than deterministic.
- Expression levels of genes are proxies for the activity level of the proteins they encode, although there are numerous examples where activation or silencing of a regulator is carried out by post-transcriptional protein modifications

Methods

- Gene expression levels, treated as continuous variables, can be assumed to follow a multivariate normal distribution, and to be consistent with a Bayesian network with linear Gaussian conditional densities.
- The prior of this network is characterised by a prior network reflecting our belief in the joint distribution of the variables in question, and equivalent sample size (ESS) effectively behaving as if it was calculated from a “prior” data set of that size. For instance, without a priori knowledge of the regulatory network, the prior network could be one where all expression levels are independent in order to avoid explicitly biasing the learning procedure to a particular edge.
- The common approach to the learning procedure starts with a training set and evaluates networks according to an asymptotically consistent scoring function that is obtained through the Bayesian framework.
- In the case of B-course software (http://b-course.hiit.fi) to be used here, discretisation of continuous data has been applied to capture the nonlinear relationship between variables and the choice of prior is such that the resulting ESS prior distribution is close to Jeffrey’s prior. The software infers causal relationship according to the statistical dependence under some additional assumptions concerning latent variables. Mathematical details, including the definition of Jeffrey’s prior, are given elsewhere
- The so-called causal structure assumes that dependencies between variables are due to causal relationships between variables in the model.

Results



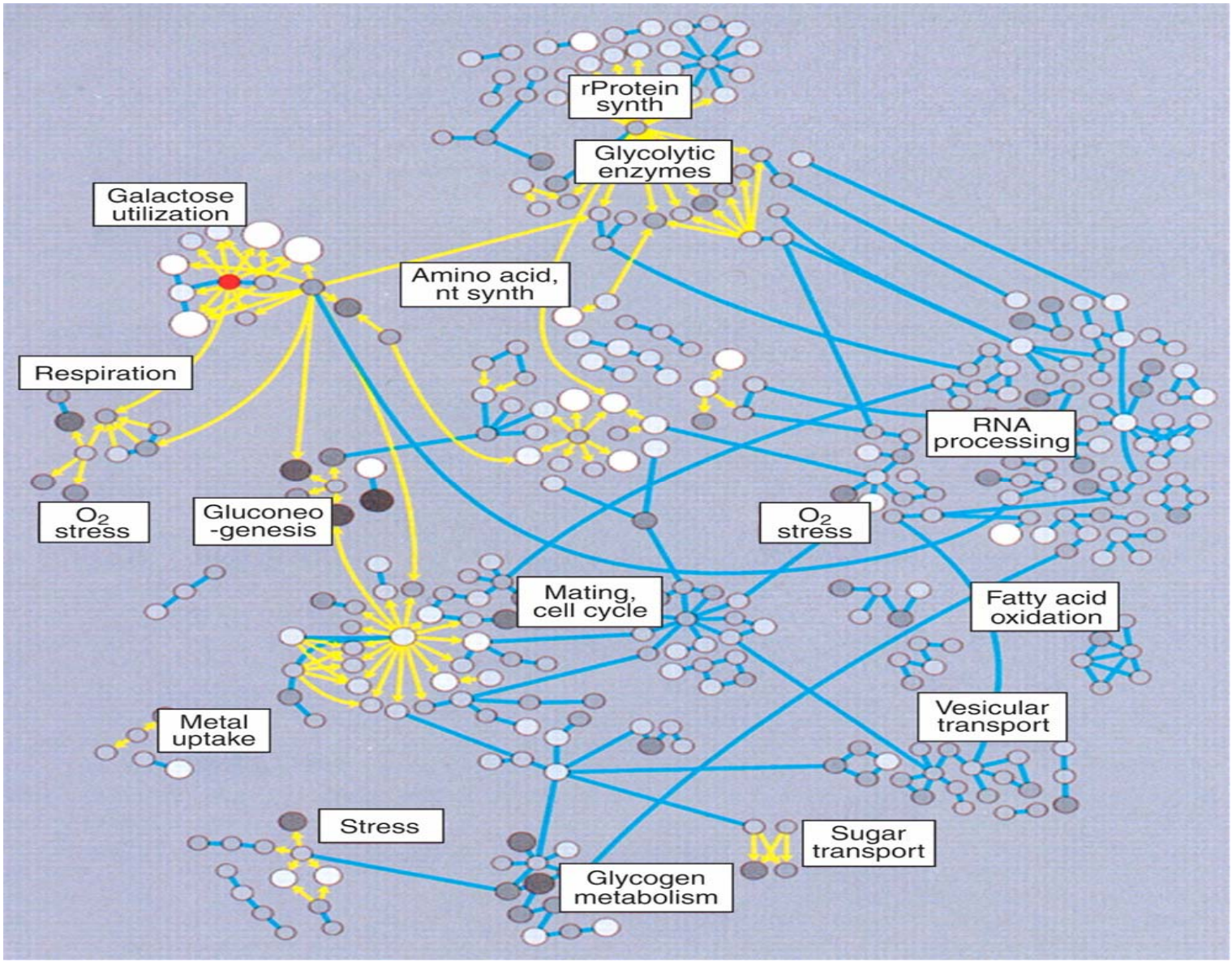
From Zhao et al. BMC Proc 1:S52, 2007, with b-course

Left. Importance of the dependencies. **Right**. Solid arc has direct causal influence (direct meaning that causal influence is not mediated by any other variable that is included in the study). Dashed arc indicates there are two possibilities, but we do not know which holds. Dashed line without any arrow heads indicates there is a dependency but we do not know the reciprocal dependence.

Summary

- The series of papers on these data stress the importance of Intermediate phenotypes. Without a priori biological hypothesis, it serves as an exploratory tool for subsequent confirmatory analysis.
- This particular analysis highlights the potential usefulness of pathway analysis but with great work to be done
- An apparent limitation of this work, though not uncommon in gene-expression studies, is the relatively small sample size used. To fully elucidate the biological pathways involved may be difficult, as for instance CYCS is involved in six pathways according to http://escience.invitrogen.com/ipath/.
- Statistical robustness and biological interpretability remain as the two main challenges for Bayesian network analyses, to which replication, bootstrap and benchmarking have been proposed.
- Our inference of gene networks also exploits the covariance structure of the data, like structural equation modelling, but is exploratory or hypothesis-generating rather than confirmatory or hypothesis-driven. A number of other software systems are of interest, e.g. ASIAN (a web-based regulatory network framework, http://eureka.cbrc.jp), deal.

Network Perturbation Model of Galactose Utilization in Yeast



Hood et al. (2004) Science

Metabolic Pathways

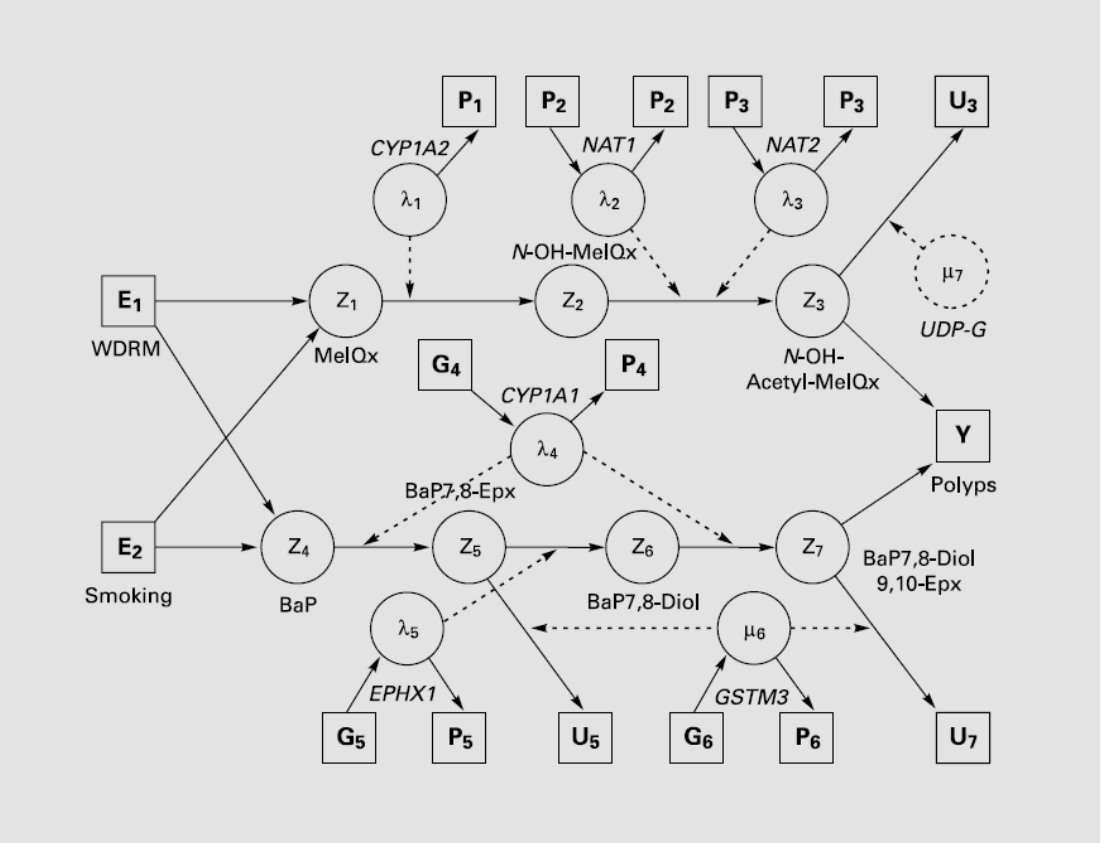


Fig. 1. Representation of the proposed metabolic pathways as a Directed Acyclic Graph (DAG). Boxes represent measured quantities for exposures (E), genotypes (G), metabolic phenotypes (P), the outcome polyps (Y), and urinary metabolites U (not included in this analysis). Circles represent the unobserved quantities, intermediate metabolites (Z) and metabolic activation (λ) and detoxification (μ) rates. Not shown are the genotypespecific population mean metabolic rates ($\overline{\lambda}$ s, $\overline{\mu}$s), interpersonal variance in these rates σ^2, and metabolic phenotype measurement error variance (ω^2); see [9] for a discussion of these elements.

Conti et al. 2004. Hum Hered. With WinBUGS. The model was linked with GAW12 data generation. This example is generic in that a hypothesis-based pathway model can be fitted by elementary assumptions.

Bayesian Networks

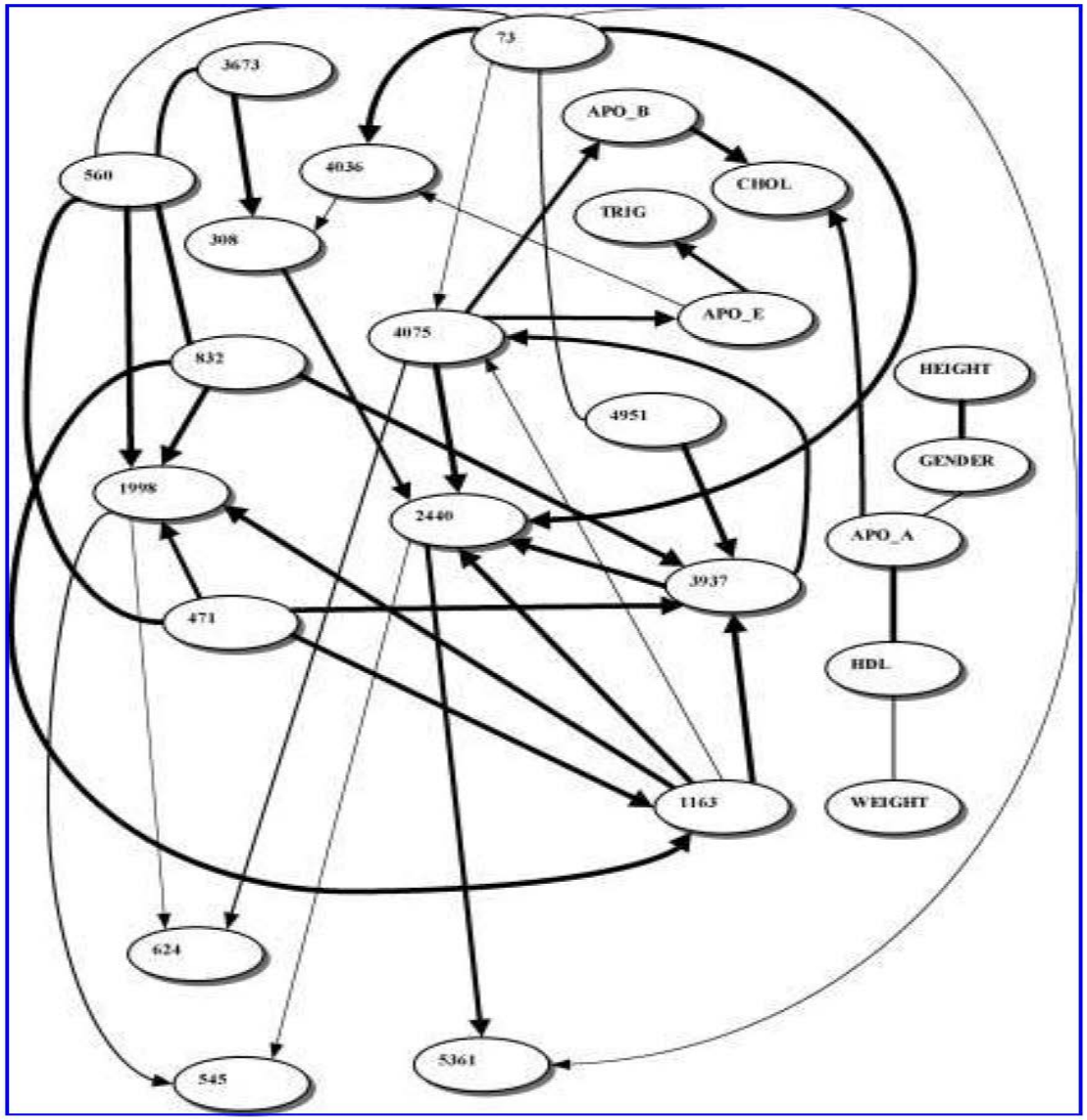


FIG. 1. Learned Belief network relating *APOE* SNPs to plasma **apoE** levels in Jackson, MS. Node legends: numbers refer to corresponding SNPs (see Fig. 1 in Nickerson *et al.* [2000] for an *APOE* SNP map). APO_E, APO_A, APO_B, TRIG, CHOL, and HDL stand for levels of apolipoproteins E, AI and B, triglycerides, cholesterol and high-density lipoprotein cholesterol, respectively. Line thickness corresponds to the relative edge strength (see Table 1.)

From Rodin et al. 2005, J Comp Biol. Exploratory Bayesian network analysis which involves both phenotype and genotype data and covariates

Health Selection in Civil Servants

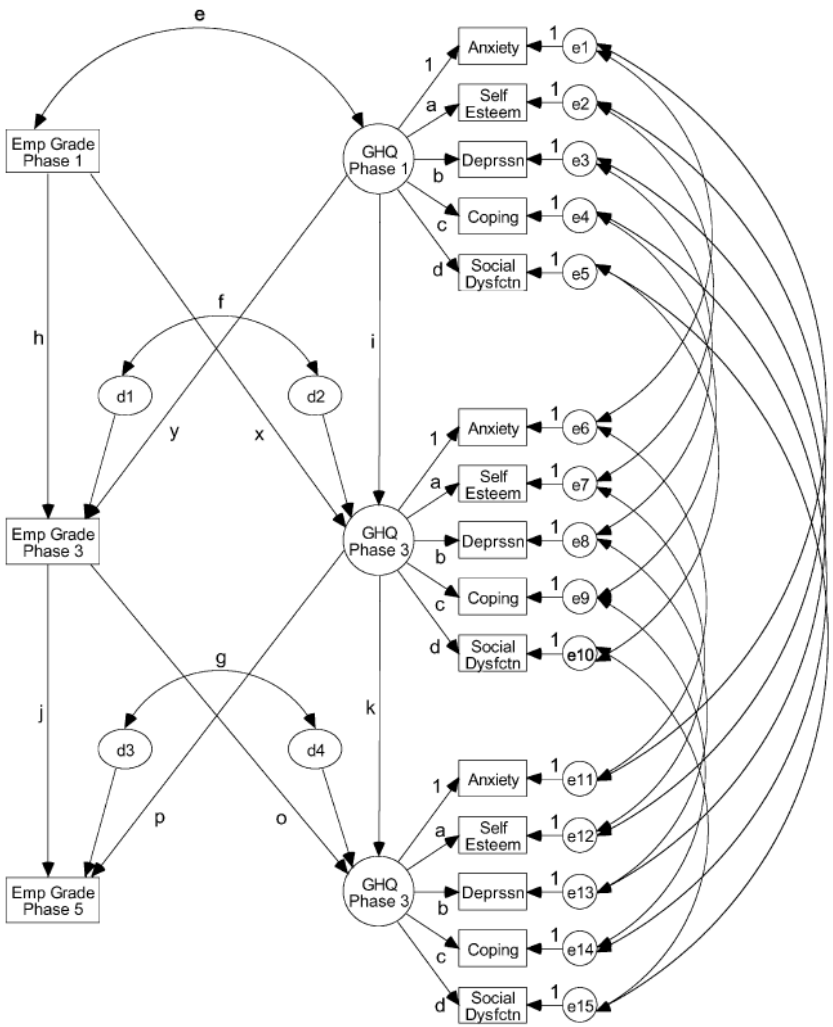


Fig. 1. Cross-lagged longitudinal analysis using structural equation models employment grade and GHQ at Whitehall phases 1, 3 and 5.

From Chandola et al. 2003, Soc Sci Med, with AMOS. It showed no selection. Similar findings were made with addition of phase 7 data.

Structural Equation Modelling

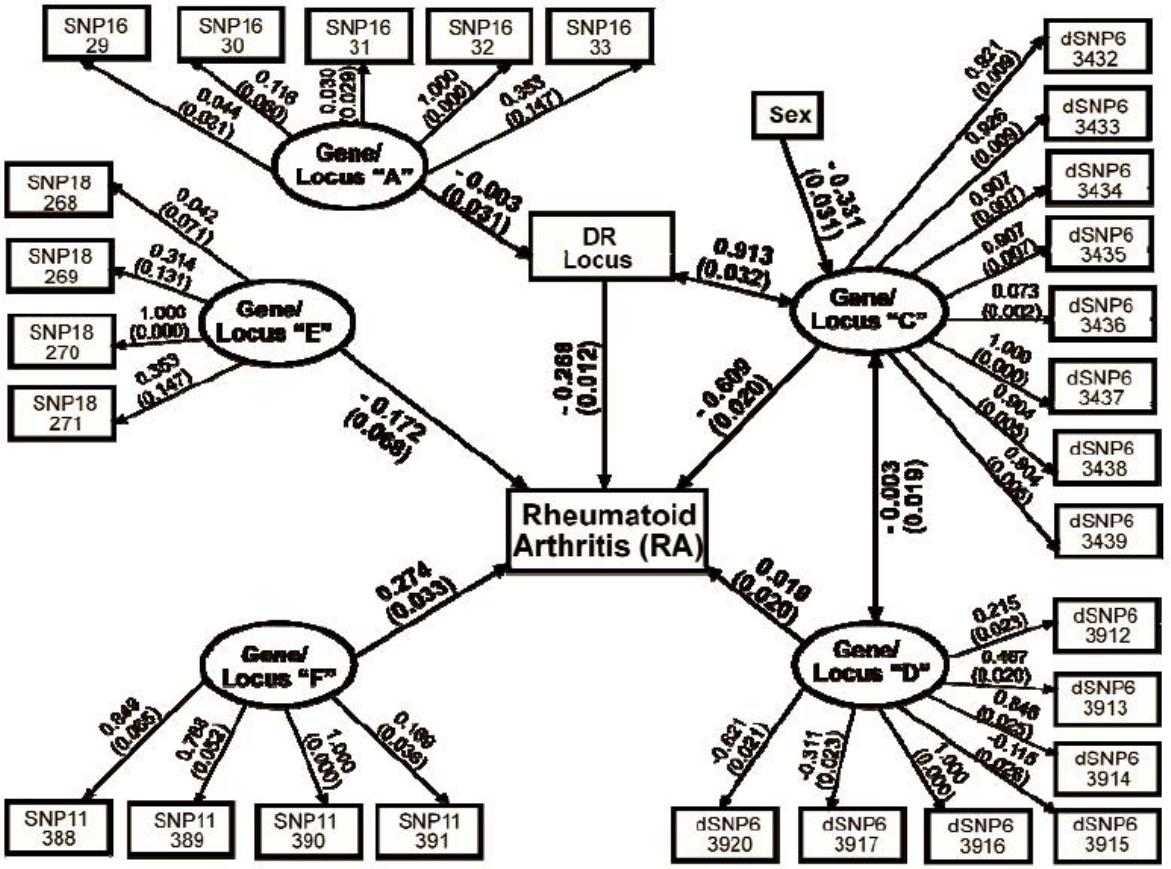
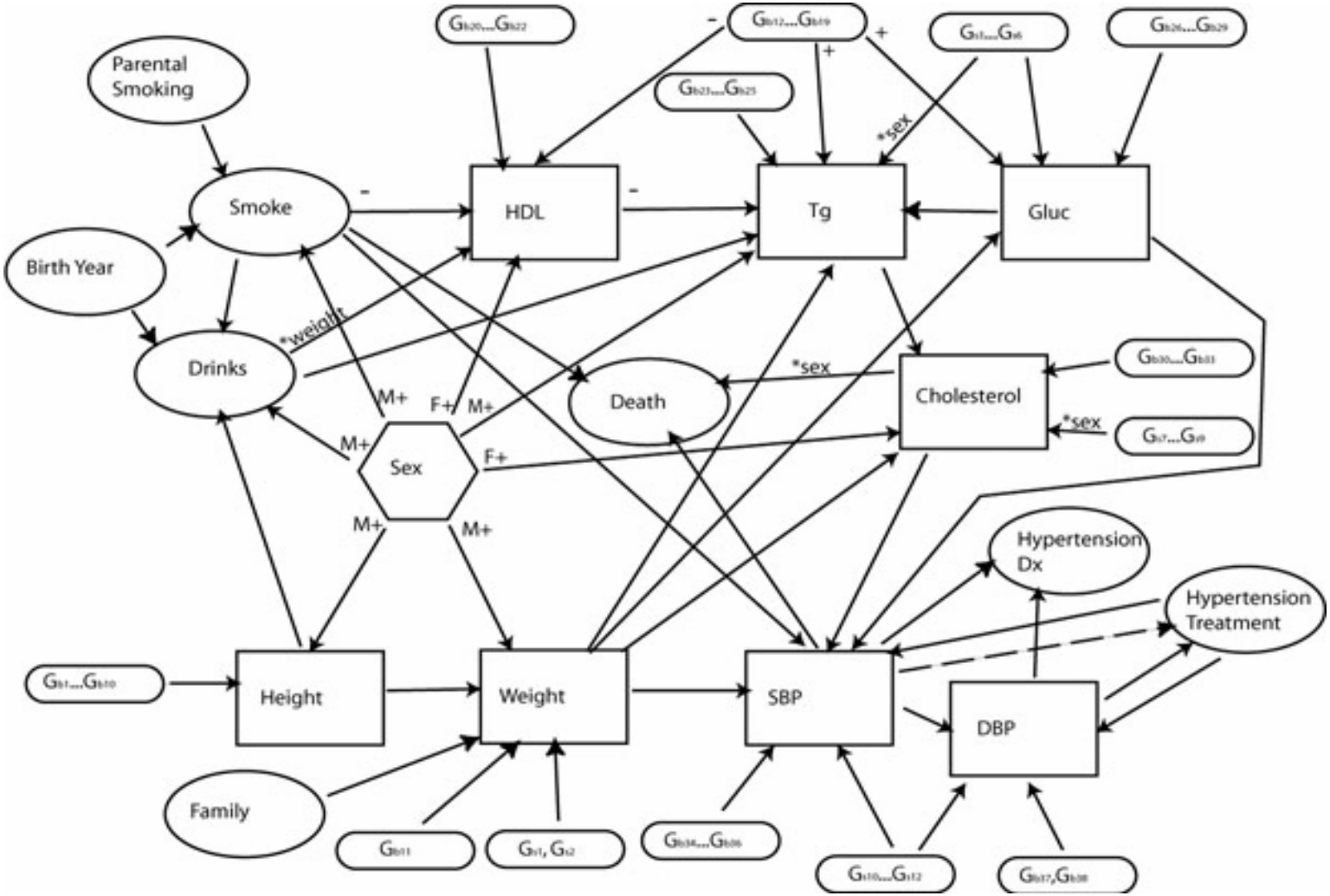


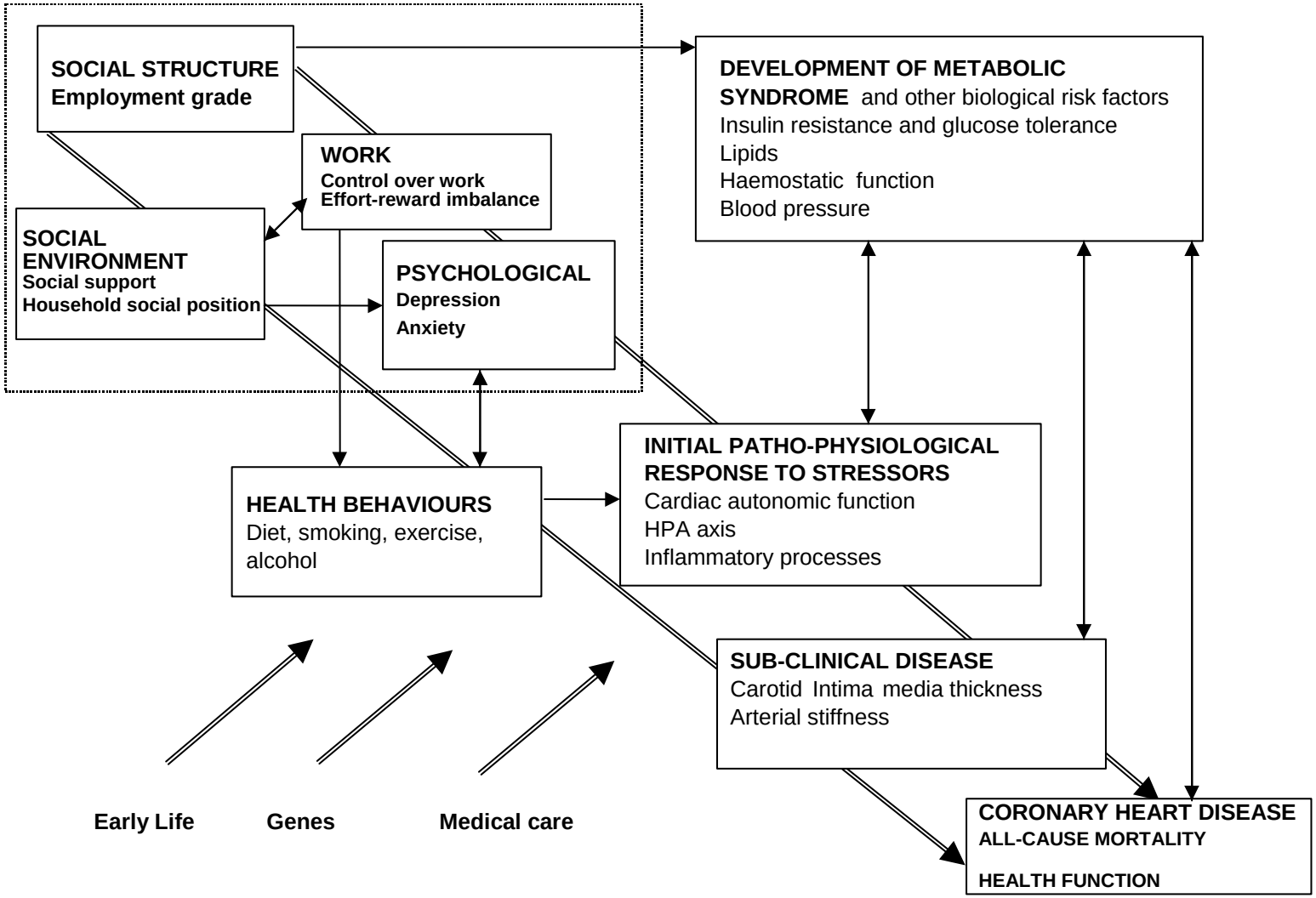
Figure 1. Evaluation of Rheumatoid Arthritis (RA) Data (Replicate #64) Using Structural Equation Modeling (SEM). Relationships in the measurement model between observed indicators (rectangles) and latent variables (ovals) are depicted by loading coefficients and standard errors in ( ) above single-headed unbolded arrows. Relationships in the structural model between latent variables and between latent and observed variables and the outcome, RA, are depicted by path coefficients and their standard errors in ( ) above single-headed bold arrows. Correlations are depicted by double-headed bold arrows.

From GAW15, with Mplus



GAW 13. Relationships between simulated traits and genes. Arrows indicate causal relationships between traits. Most correlations are positive, but a "-" indicates a negative correlation. An "*" and trait name next to an arrow indicates that the relationship is mediated by the named trait. Daw *et al. BMC Genetics* 2003 **4**(Suppl 1):S3

Whitehall II Study of 10,308 British Civil Servants from 1985



Master plan of research activities in the study

A Call

“A strength of the package is that it aims to be the seed point for a general compilation of such methods, which means it implements many different kinds of methods, not just the ones developed by the author'', the author would like to take this opportunity to invite comments, suggestions and contributions to consolidate the package.

J Stat Soft v23 n8

Acknowledgement

- useR!2008 organising committee
- Colleagues for inspiring work
- Adapted from a number of presentations by myself and colleagues (Chris Cannings, Michael Stumpf, Jonathan Marchini, David Clayton, Heather Cordell, Elizabeth Thompson, Joe Terwilliger)

Exercises

- Hardy-Weinberg equilibrium (genetics, hwe)
- Linkage disequilibrium (genetics)
- Single-point analysis (glm, qqplot)
- Multipoint analysis (haplo.stats)
- Risk calculation (BayesMendel)
- HapMap data (snpMatrix)
- GWA analysis (GenABEL, SNPassoc)
- MCMC (WinBUGS, R2WinBUGS)
- Meta-analysis (meta)

Bayesian Methods

$$B(n, p) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}, \quad f(p) = \frac{p^{v-1}(1-p)^{w-1}}{Be(v, w)},$$

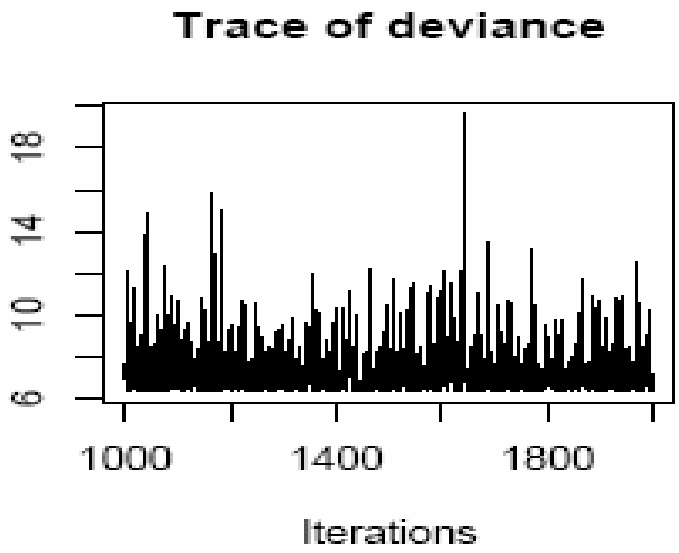
$$p(p \mid k, n) = \frac{p^{k+v-1}(1-p)^{n-k+w-1}}{Be(k+v, n-k+w)}$$

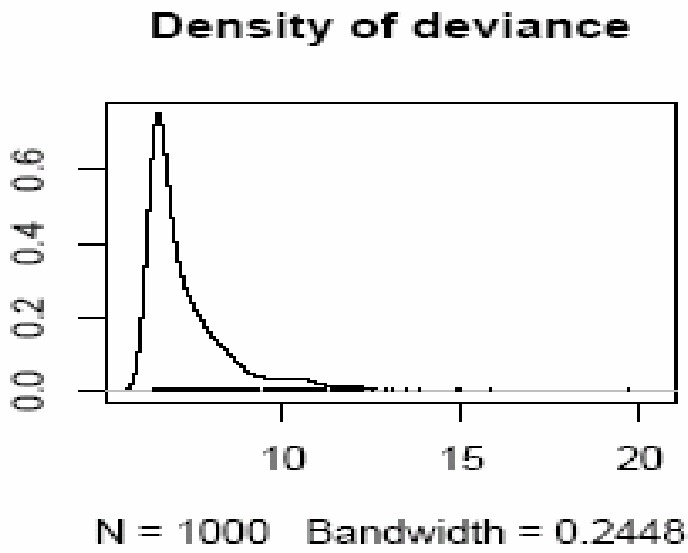
```
# bs-model.txt:
model {
# likelihood
   k ~ dbin(p, n)
# prior
   p ~ dbeta(nu, omega)
}
```

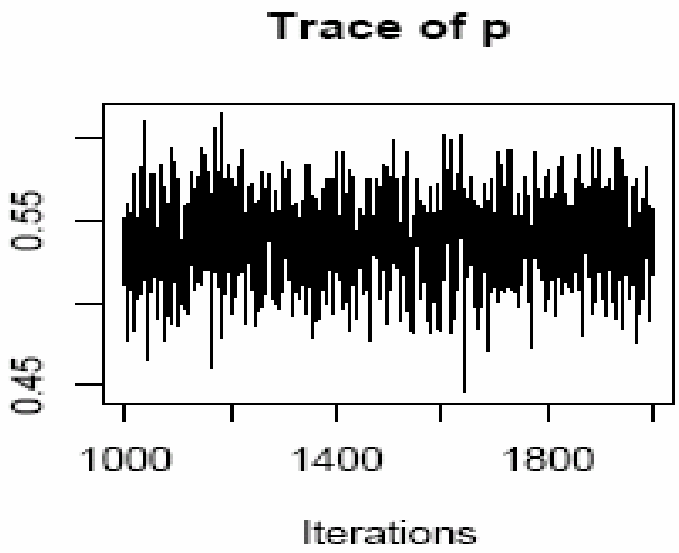
R2WinBUGS

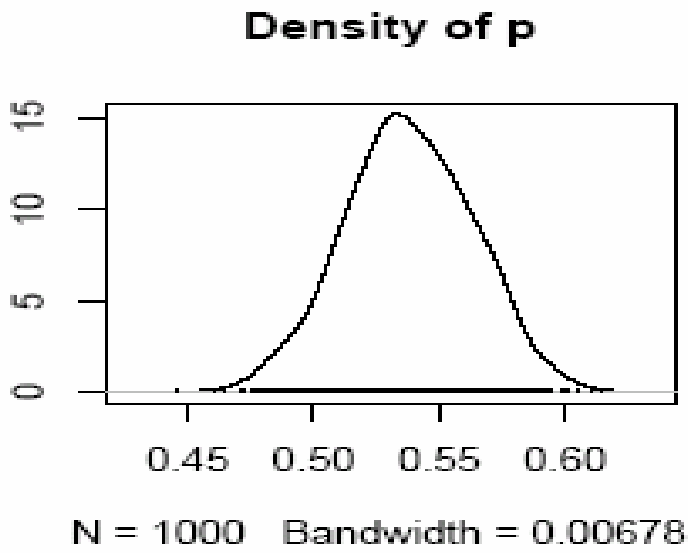
```
library(R2WinBUGS)
s <- list(k = 201, n = 372, nu = 1, omega = 1)
parms <- "p“
inits <- function() list(p=0.5)
s.bugs <- bugs(s,inits,parms,"bs-model.txt", n.chains=1,
   n.burnin=1000, n.iter=11000)
> s.bugs
mean  sd 2.5% 25% 50% 75% 97.5%
p         0.5 0.0  0.5 0.5 0.5 0.6   0.6
deviance  7.3 1.4  6.4 6.5 6.8 7.7  11.0
pD = 0.9 and DIC = 8.3
plot(s.bugs)
```

MCMC Plot for Bionomial Sampling









Meta-analysis: an Example

```
> attach(toplist)
> library(meta)
> by(toplist,snp,function(x) metagen(b,se,data=x))
snp: rs1000587
                  95%-CI %W(fixed) %W(random)
1 -0.0584  [-0.1176;  0.0007]    26.13      23.44
2 -0.1044  [-0.1790; -0.0297]    16.41      18.11
3 -0.0002  [-0.0729;  0.0725]    17.29      18.70
4  0.0100  [-0.0488;  0.0688]    26.44      23.57
5 -0.0366  [-0.1182;  0.0450]    13.73      16.18

                                        95%-CI         z        p.value
Fixed effects model      -0.0348  [-0.0651; -0.0046] -2.2575   0.024
Random effects model  -0.0362  [-0.0769;  0.0044] -1.7462   0.0808
Quantifying heterogeneity:
tau^2 = 9e-04; H = 1.33 [1; 2.19]; I^2 = 43.3% [0%; 79.2%]
Test of heterogeneity:
    Q d.f. p.value
 7.05    4   0.133
```